Assessing cultural influences on cognitive test performance

Helms-Lorenz, M.

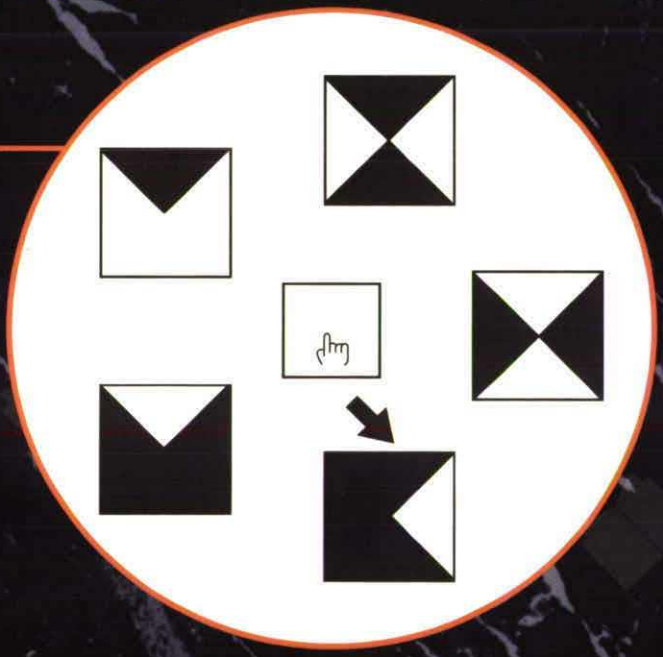
Publication date:
2001

Document Version
Publisher's PDF, also known as Version of record

DISSERTATION

SOCIAL & BEHAVIORAL SCIENCES

Assessing Cultural Influences on Cognitive Test Performance: A Study with Migrant Children in the Netherlands



Michelle Helms-Lorenz

ASSESSING CULTURAL INFLUENCES ON COGNITIVE TEST PERFORMANCE: A STUDY WITH MIGRANT CHILDREN IN THE NETHERLANDS

ASSESSING CULTURAL INFLUENCES ON COGNITIVE TEST PERFORMANCE: A STUDY WITH MIGRANT CHILDREN IN THE NETHERLANDS

PROEFSCHRIFT

ter verkrijging van de graad van doctor
aan de Katholieke Universiteit Brabant,
op gezag van de rector magnificus, Prof.dr. F.A. van der Duyn Schouten,
in het openbaar te verdedigen ten overstaan van
een door het college voor promoties aangewezen commissie
in de aula van de Universiteit op vrijdag 27 april 2001 om 13.45 uur

door

Michelle Helms-Lorenz
geboren op 17 mei 1966
te Johannesburg Zuid-Afrika

Tilburg University

Promotoren: Prof.dr. Fons van de Vijver
Prof.dr. Ype Poortinga

The publication of this thesis was funded by the J. E. Jurriaanse Stichting, Rotterdam.

ISBN 90-75001-37-1

Acknowledgements

This thesis would not have been completed without the help and patients of a number of people. Most importantly I would like to express my gratitude and appreciation towards my supervisor, dr. F. van de Vijver, for patiently hanging in there for so long. Fons, your stamina was more than mine, and for this I am grateful. It was a privilege to have been given the opportunity to witness your expertise. Ype Poortinga, you I would like to thank for granting me this opportunity, for believing in me and for encouraging me in dark hours. I owe the university (in particular Jan Boelhouwer and later Ton Heijnen) a special word of thanks for supplying me with all possible facilities, enabling me to work from home. John van der Beesen and Ton Aalbers deserve a special word of thanks for their kindness and help in the making (or should I rather say baking?) of TAART. A number of students were involved in my data collection. I would like to thank you all for your contributions: Marijn van Dijk, Ellen Meyer, Lucas Lichtenberg, Astrid Voogd, Tessa van der Sluis, Martine Westenbroek, Tamara Rijkers and 5 students of the University of Groningen. I would also like to thank my collegues Corine, Judit, and Tina for breaking the solitude that this kind of project brings along. Thanks for the divergent conversations, for your interest and encouragement throughout the years. I have numerous people to thank for looking after my children so that I could work. A few I would like to thank here: the deceased Silvia de Jong, Wies de Groot, and Jadga. As the list of acknowlegement grows, it becomes more personal. My parents in-law Leida and Evert Helms deserve special gratitude for their contribution during the last phase of my project. My friend Hellen, I thank you for your prayers, your unselfish help and for your warm involvement in my life. My sons Bernie, Paulie and Henrie, to whom I devote this thesis, I thank for the sacrifices made for me to finish what I had begun before you were born. And then Jan, how can I thank you for giving me space and time when it was scarce? To you I am more than grateful. Finally, without the blessing of our Father in Heaven none of this would have been possible.

Contents

De grenzen van mijn taal zijn
de grenzen van mijn wereld.
Ludwig Witgenstein

Prologue

Humans differ in their cognitive abilities. We differ in the way we solve everyday problems, our ability to understand complex ideas, our ability to reason, and the time we need to make complex decisions. For more than a century many researchers (e.g., Binet & Simon, 1916; Carroll, 1993; Galton, 1883; Jensen, 1985; Spearman, 1927) have been trying to unravel the nature of intelligence. In pursuit of a theory, many approaches have been tried and rejected (Irvine & Berry 1988). In the following section a brief overview of some of the main approaches will be given. The aim is to indicate the position of the present study within the domain of intelligence theory. The second aim of the overview is to illustrate that none of these approaches pays attention to what Irvine and Berry call *the law of cultural differentiation*. This law can also be referred to as *Ferguson's law*, as he was the first to formulate the law:

> Cultural factors prescribe what shall be learned and at what age: consequently different cultural environments lead to the development of different patterns of ability (Ferguson, 1956, p. 121).

Approaches to Human Cognitive Functioning

Psychometric Approach This approach is characterized by explorative statistical analyses of test responses. It lacks a substantive definition of intelligent behavior. The simplest definition of intelligence put forward in this tradition is: intelligence is what intelligence tests measure (Boring, 1923). In this bottom-up approach, test batteries determine the scope of the theory.

The development of statistical techniques, such as Pearson's Correlation Coefficient, and Factor Analysis led to a number of psychometric discoveries; such as the observation of the 'positive manifold' phenomenon (Spearman, 1927). This refers to the repeated finding of positive correlations between test results obtained with tests for different abilities. Spearman explained this phenomenon by postulating a general intelligence factor (*g* factor). The g factor represents what all (valid) cognitive tests have in common. This first model used to explain human abilities has remained influential to this day; for example, later hierarchical models were based on Spearman's *g* (e.g., Gustafsson, 1984).

Researchers like Thurstone (1938) found evidence for specific, uncorrelated factors, incompatible with the notion of a general intelligence factor. Specific group factors such as memory, verbal comprehension, and number facility were found to form specific profiles for individuals.

Such non-hierarchical models seem more incompatible with the hierarchical models of Jensen and others than they actually are. Factor analytic (rotation) methods and heterogeneity of samples have been argued to be responsible for the differences found between the two kinds of models. An orthogonal rotation such as VARIMAX will "rotate the general factor away" (e.g., Gustafsson, 1984).

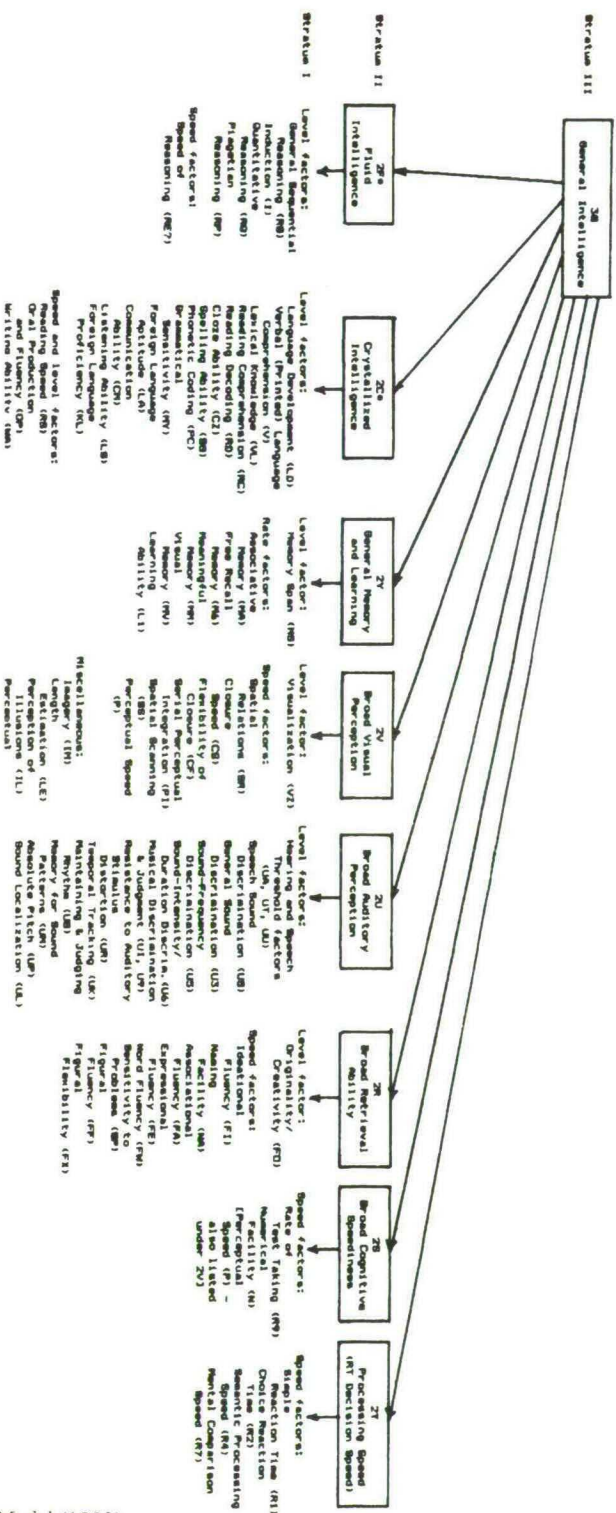


Figure 1. Carroll's Model (1993)

Carroll (1993) reanalyzed 460 data sets obtained between 1927 and 1987. The model Carroll proposed, is a hierarchy comprising three strata: Stratum I includes narrow, specific abilities (e.g., spelling ability): Stratum II includes group factors (e.g., fluid intelligence, crystallized intelligence); and Stratum III, represents a single general intelligence factor. The Stratum II group factors have different relationships with the *g* factor (Stratum III). In figure 1 an overview is given of his model. The distance between the g-factor and Stratum II factors provides an approximate indication of the strength of their relationship.

Biological Approach This approach is historically the oldest and dates back to Galton's (1883) account of intelligence in terms of psychophysical abilities (such as strength of handgrip or visual acuity).

Since the 1980s advances in technology have enhanced the quality and quantity of studies seeking a biological basis of intelligence. Hendrickson and Hendrickson (1980) proposed the "string length" measure of averaged evoked potentials (AEPs) to be a physiological manifestation of intelligence. Now this approach includes measures related to electroencephalography, cortical neurons (Ceci, 1990), cerebral glucose metabolism (Haier, 1993), evoked potentials (Caryl, 1994), nerve conduction velocity (Reed & Jensen, 1992), sex hormones and others (cf. Neisser et al., 1996).

Researchers in this field are interested in aspects of brain anatomy and physiology that are potentially relevant to intelligence. All the measures have a common purpose: finding a biological basis for intelligence. Melis (1997) summarizes the findings of this approach and concludes that associations have been found (despite inconsistencies) between brain functions and IQ measures, but the main problem is that the mechanisms behind these established relationships remain unknown.

Cognitive Approach This approach is often referred to as the information processing approach. Irvine and Berry (1988) group the cognitive and the biological approaches together in a single category. Hunt, Frost, and Lunneborg (1973) introduced the cognitive-correlates approach, whereby scores on laboratory tasks of cognition were correlated with scores on psychometric intelligence tests. A prototypical example of information processing is the inspection time (IT) task (Nettlebeck, 1982), in which two vertical lines are briefly presented tachistoscopically, followed by a visual mask. The two lines differ in length. The subject's task is to indicate which line is longer. Correlations between this task and traditional measures of IQ appear to be about .4.

Sternberg (1977) introduced the idea of cognitive components, in which the performance on complex psychometric tasks was decomposed into elementary information-processing components. The focus is on defining the information content of reasoning in terms of stages of processing. Each one of these can be defined, and its latency used as a measure of the process (cf. Irvine & Berry, 1988). Tasks are described in terms of cognitive elements or steps that are successively involved in problem solving. Stages

that have been isolated include: an encoding phase, an inference phase, a mapping phase, an application phase and an optional justification phase.

Sonke (2001) carried out a cross-cultural study that combines physiological measures with the information processing tradition. One of the aims of her study was to find similarities between reaction time (RT) patterns using elementary cognitive tasks (ECT) and patterns of concurrent Event Related Potentials (ERPs). This method of inquiry is pioneering in view of the fact that an attempt was made to localize the information processing stage that is responsible for cross-cultural differences in RTs in terms of brain processing parameters. Despite the absence of the hypothesized ERP effects, this kind of approach offers promising perspectives to investigate cross-cultural differences on ECTs in closer detail. The complexity of this kind of research should not discourage future attempts of refining it.

Developmental Approaches The ontogenetic development of cognitive functions has been described, among others by Piaget (1947), Vygotsky (1962), and Fischer (1980) who formulated the Skill Theory. The essential distinction of Piaget's original approach is the assumption that development is programmed in stages. The exact time at which an individual reaches a stage is of minor importance compared to the fixed progress of successive stages supposedly found in all humans during their cognitive development, and the final stage that everyone does reach eventually. It was only later that the final developmental stage was argued not to be reached by some cultural groups (Dasen, 1972). In the present study the developmental aspect of intelligence is not of primary concern, although one of the theories, the Skill Theory, was used to determine task complexity (cf. Chapter 2).

Cross-Cultural Studies

Irvine and Berry (1988) state that a "theory that does not encompass cross-cultural empiricism has no apriori claim to universality. By definition, from the law of cultural differentiation, such theory can expect confirmation only within its own culture, because it is equipped for that purpose and no other" (p. 7). To illustrate this point Irvine and Berry (1988) give a detailed review of cross-cultural research that attempts to test the universality of "western facts". Born (1984) reveals cultural inconsistencies in the direction of sex differences in performance. Her painstaking study revealed that conclusions concerning sex differences in performance do not show the cross-cultural stability that seems to be implied by western psychologists. Lloyd and Pidgeon (1961) showed that different cultural groups show dissimilar practice effects when tests are administered repeatedly. For more readings on the abundant cross-cultural studies that often are in disagreement with western empirical findings the reader is referred to Irvine and Berry (1988).

The approaches discussed in the previous section, as well as the need for culture-informed analysis, emphasized in this section, enhance our understanding of human

abilities. The application of the theories in practical settings requires careful operationalization of constructs. And, to use Irvine's (1979) words: "to lay claim to validity, constructs must, in turn, become part of a more encompassing scientific statement that will allow the prediction of future events from observations made in the past, by linking these events mathematically" (p. 301). In the following section this aspect will be considered.

The Validity of Instruments in Multicultural Settings

Within the different approaches to human ability, all kinds of cognitive instruments have been applied. The success and frequent use of tests can be attributed to their good predictive validity of external criteria, notably school and job success. Test scores correlate with the desired behavior needed for real life situations.

In multicultural settings cognitive tests reveal a consistent finding. Minority group members and individuals not coming from the western world, perform less well on these tests, compared to their western counterparts (for example, in the US Blacks score on average 1.0 SD below Whites on IQ batteries). This raises a fundamental question: Are these differences in performance "real" or the result of test bias? In cross-cultural research test bias is defined as "all nuisance factors threatening the validity of cross-cultural comparisons" (Van de Vijver & Leung, 1997a, p. 10).

In order to understand the nature of cross-cultural differences in performances on cognitive tests, we need to investigate what tests are measuring. If cognitive ability tests are reflecting additional factors besides intelligence, then the nomenclature of tests should be expanded and references to the kinds of ability investigated need to be defined more sharply. Cognitive ability tests may well need to be referred to with a distinct term in this sense, such as "context-related" intelligence tests. To illustrate the confusion that arises when a single construct is used to refer to different levels of human functioning, the following example can be used. Obesity is a combined measure of both body weight and height (it is a combined measure of more than one body parameter). Obesity cannot be derived from a measure of a person's weight alone. Weight can be an indicator of obesity only, if a person's height is also known. Actual computations can be made when the mathematical formula of the relationships between the three parameters is known. In the same sense, if cognitive tests are actually measuring cognitive functioning combined with some other factor(s), then (the sizes and relationships between) these other factors need to be known, before cognitive ability can be determined.

In the case of intelligence, confusion can be said to result from the fact that we are dealing with a confounded construct that has the same name as one of its components. The reason for this confusion is that the culture parameter may well be constant within a (homogeneous) group, therefore not influencing computations for members of the same group. As soon as the 'culture' factor is not constant (across groups) then the determination of intelligence is seriously jeopardized if culture's impact is ignored.

This kind of reasoning is also reflected in the concern that is frequently expressed

in the cross-cultural literature about the non-equivalence or bias of test scores. In recent years a distinction between three categories of inequivalence has gained prominence (Van de Vijver & Leung, 1997), viz: *Construct bias* occurs when the construct measured is not identical across cultural groups. *Method bias* is a generic name for all sources of cross-cultural score differences deriving from the characteristics of a test (e.g., stimulus familiarity), samples (e.g., differential education or motivation), or administration. *Item bias* or Differential Item Functioning (DIF) refers to measurement artifacts at item level. DIF, in the psychometric sense, occurs "if individuals with equal ability but from different groups do not have the same probability of answering an item correctly" (Shepard, Camilli, & Averill, 1981, p. 319).

One of the early researchers to recognize the joint influence of genetic and environmental facets in human abilities was Cattell (1940) who developed the notion of fluid and crystallized intelligence. Whereas fluid intelligence refers to genetic endowment, crystallized intelligence is the product of formal schooling, socialization (Child, 1954) and experience. The development of Cattell's Culture Fair Intelligence Test (Cattell & Cattell, 1963) and Raven's Progressive Matrices (Raven, 1938) can be seen as attempts to create instruments that measure pure 'g'. Unfortunately these tests did not lead to the desired outcome. Vernon (1969) and Anastasi (1976) provide ample evidence that the performance on fluid intelligence tasks is not free of cultural influences (e.g., socioeconomic status).

Hebb (1949) distinguished between two types of intelligence: Intelligence A (innate potential, or biologically determined ability), and B (the functioning of the brain as a result of actual development, or environmental influence). Vernon (1979) added the notion of Intelligence C; it is distinguished from the other two, as intelligence measured by conventional psychometric tests. The important differentiation of Intelligence C underlines the typical pitfall of test application in cross-cultural settings. Cognitive instruments are too readily assumed to be measuring intelligence or "IQ" only.

Current Study

The focus of the present project is not so much to enhance awareness of the value of cross-cultural cognitive research in general, but rather to apply cross-cultural findings to some aspects of the psychometric approach, using a set of instruments developed under the information-processing approach, in order to determine their validity in a culturally heterogeneous society.

The population in the Netherlands has changed from a fairly homogeneous cultural group to a heterogeneous one over the last four decades. The influx of migrant workers (mainly from Morocco and Turkey) after the World War II, was encouraged by the government and led to fast restoration of the damages caused by World War II, as well as to further economic growth. The entry of family members of these workers led to a large group of second and third generation of citizens with Turkish and Moroccan roots. Citizens from the former Dutch colonies (Indonesia, Dutch Antilles, and

Surinam) form another part of the migrant population. In addition, recent decades have seen many refugees enter the country from all over the world.

This diverse multi-cultural population has formed a new challenge for the educational system in The Netherlands. Developments, such as the right of tuition, for a certain number of hours per week, in the own language and culture have been implemented; Islamic schools have been opened, etc. This new educational setting calls for scrutiny of cognitive tests commonly used in the Netherlands. High quality decisions based on test results are necessary and essential for the functioning of a healthy and just society.

Examinations of the suitability of ability tests in multi-cultural applications was stimulated by repeated findings that subject- and instrument-related factors negatively influence cognitive performances of minority group members. In Chapter 1 this point is elaborated; it supplies guidelines to avoid typical pitfalls in multi-cultural assessment. A description is given of types of biases that can threaten cross-cultural comparisons of test scores. An overview is given of sources of bias and of subject-related factors differentially influencing test performance. The last part of the review describes ways to increase suitability of tests in multi-cultural settings.

A series of cognitive computerized reaction time tasks called TAART (an acronym for Tilburgse Allochtone en Autochtone Reactietijd Test) was used for the present project. TAART consists of so-called elementary cognitive tasks (ECT) (e.g., Vernon, 1987). The focus is on speed (reaction time) rather than on accuracy. The tasks are simple enough for all subjects to respond correctly to all items. Successive tasks show increasing cognitive complexity. In developing this instrument the most important objective was to reduce the influence of potentially biasing subject-related factors on test performance, such as verbal skills, cultural knowledge, and test-wiseness. The test is virtually non-verbal. The interaction between the tester and the testee is reduced as much as possible, and the role of the tester in test administration is peripheral, compared to his/her role in the administration of traditional paper-and-pencil tests. Furthermore, it was attempted to reduce instrument-related biasing factors, such as the cultural loading of test items. The stimulus material consists of geometric figures. The assumption underlying the choice of geometric stimuli, is that cultural groups show fewer differences in familiarity when stimuli are less culturally loaded (entrenched). Ample opportunity for practice is given in order for the subjects to become familiar with the stimulus material and the test setting.

The test's theoretical foundation can be traced back to the cognitive approach of human ability. In 1868 Donders was the first to start research measuring the time needed to perform cognitive tasks (Carlson & Widaman, 1987). In 1883 Galton related reaction time performance to intelligence. The development and success of complex intelligence tests (e.g., Binet & Simon, 1916) diminished the interest in reaction time measures. Hick (1952) gave a new impetus to this research by showing that mean reaction time increased linearly with the logarithm of the number of possible response choices. The notion that individual differences in intelligence can be traced back to differences

in speed of information processing led to the use of reaction time tests in intelligence research (e.g., Jensen, 1987; Jensen & Munro, 1979).

A major reason for our choice to develop and validate an instrument measuring reaction time, was the repeated finding that performances on elementary reaction time tasks are indeed correlated, albeit moderately, with intelligence. As tasks become more complex (as in choice reaction time tasks) the correlation can become as high as -.30 or -.40 (e.g., Jensen, 1987). This led Vernon (1983) to suggest that reaction time is related to basic cognitive operations involved in many forms of intellectual behavior. He elaborated this by adding that individual differences in intelligence can be attributed, to a moderate extent, to variance in speed or efficiency with which individuals can execute these operations.

For this study TAART was applied to a sample of 1,462 subjects including migrants and majority group members, aged 6-12 years, in The Netherlands.

Chapter 2 deals with the first aim of this project, namely to investigate Spearman's Hypothesis (SH) using a fairly large number of culture-reduced tests administered to a multi-cultural sample of school children in The Netherlands. This line of research investigates the plausibility of the hypothesis stating that performance differences found between cultural groups are due to real cognitive ability differences. The hypothesis can be traced to the psychometric approach mentioned earlier on. The hypothesis, put forward by Jensen (1985), is based on Spearman's observation in 1927 that performance differences between cultural groups increase, as tasks become more complex. Jensen operationalized task complexity in factor analytic terms, as being the factor loadings on the first factor (called the tests *g loading*). Now, SH states that performance differences between cultural groups on cognitive tests increase with *g loading*. In this project the operationalization of complexity in terms of *g loading* to test SH is investigated. In our analysis an attempt is made to decompose *g* in verbal—cultural aspects and cognitive complexity. The relative contribution of complexity and verbal—cultural factors to observed cross-cultural performance differences, is compared. The choice to use culture-reduced tests, to investigate SH, minimizes the possibility of bias factors.

The second aim of this study was to investigate the role of construct-, method-, and item bias on test performances assessed with culture-reduced tests (TAART, RAKIT, and SONR), and school- achievement measures in the Netherlands. This analysis is presented in Chapter 3. As will be seen in this chapter, individual tests have been inspected for bias frequently, but testing different tests simultaneously within a single sample has hardly ever occurred in the past. Additionally, the three mentioned types of bias were investigated at the same time. Previous studies have focused either on construct or item bias, suggesting that the test investigated can be labeled 'suitable' for multi-cultural use, if these forms can be ruled out. The validity of this assumption is questioned in the last study.

References

Anastasi, A. (1976). *Psychological testing* (4th ed.). New York: Macmillan.

Binet, A., & Simon, T. (1916). *The development of intelligence in children.* Transl. E.S. Kite. Baltimore: Williams Wilkins.

Boring, G. G. (1923). Intelligence as the tests test it. *New Republic*, June 6, pp. 35-37.

Born, M. (1984). *Cross-cultural comparison of sex-related differences in intelligence tests: A meta-analysis.* Unpublished doctoral dissertation, Free University Amsterdam.

Carlson, J. S., & Widaman, K. F. (1987). Elementary cognitive correlates of g: Progress and prospects. In Vernon, P. A. (Ed.). *Speed of information-processing and intelligence* (pp. 69-100). Norwood, NJ, Ablex Publishing Corporation.

Carroll, J. B. (1993). *Human cognitive abilities. A survey of factor-analytic studies.* Cambridge: Cambridge University Press.

Caryl, P. G. (1994). Early event related potentials correlate with inspection time and intelligence. *Intelligence, 18,* 15-46.

Cattell, R. B. (1940). A culture-free intelligence test. *International Journal of Educational Psychology, 31,* 176-199.

Cattell, R. B., & Cattell, A. K. S. (1963). *Culture fair intelligence test.* Champaign, IL: Institute for Personality and Ability Testing.

Ceci, S. J. (1990). *On intelligence...more or less.* Englewood Cliffs, NJ: Prentice Hall.

Child, I. L. (1954). Socialization. In G. Lindzey (Ed.), *Handbook of social psychology* (Vol. 2, pp. 655-692). Cambridge, MA: Addison-Wesley.

Dasen, P. R. (1972). Cross-cultural Piagetian research: A summary. *Journal of Cross-Cultural Psychology, 3,* 23-39.

Ferguson, G. A. (1956). On transfer and the abilities of man. *Canadian Journal of Psychology, 10,* 121-131.

Fischer, K. W. (1980). A theory of cognitive development: the control and construction of hierarchies of skills. *Psychological Review, 87,* 477-531.

Galton, F. (1883). *Inquiry into human faculty and its development.* London: MacMillan.

Gustafsson, J-E. (1984). A unifying model for the structure of intellectual abilities. *Intelligence, 8,* 179-203.

Haier, R. J. (1993). Cerebral glucose metabolism and intelligence. In P.A. Vernon (Ed.), *Biological approaches to the study of human intelligence* (pp. 317-332). Norwood, NJ: Ablex.

Hebb, D. O. (1949). *The organization of behaviour.* New York: Wiley.Hendrickson, D. E., & Hendrickson, A. E. (1980). The biological basis of individual differences. *Personality and Individual Differences, 1,* 3-33.

Hick, W. E. (1952). On the rate of gain of information. *Quarterly Journal of Experimental Psychology, 4,* 11-26.

Hunt, E. B., Frost, N., & Lunneborg, C. (1973). Individual differences in cognition: a new approach to intelligence. In G. Bower (Ed.), *The psychology of learning and motivation* (pp. 87-122). New York: Academic Press.

Irvine, S. H. (1979). The place of factor analysis in cross-cultural methodology and its contribution to cognitive theory. In L. Eckensberger, W. Lonner, & Y. H. Poortinga (Eds.), *Cross-cultural contributions to psychology (pp. 300-343).* Lisse, the Netherlands: Swets & Zeitlinger.

Irvine, S. H., & Berry, J. W. (1988). The abilities of mankind: a reevaluation. In S. H. Irvine and J. W. Berry (Eds.), *Human abilities in cultural context.* Cambridge: Cambridge University Press.

Jensen, A. R. (1980). *Bias in mental testing.* New York: Free Press.

Jensen, A. R. (1985). The nature of Black—White difference on various psychometric tests: Spearman's hypothesis. *Behavioral and Brain Sciences, 8,* 193-263.

Jensen, A. R. (1987). Individual differences in the Hick Diagram. In P. A. Vernon (Ed.), *Speed of information processing and intelligence* (pp. 101-175). Norwood, NJ: Ablex.

Jensen, A. R., & Munro, E. (1979). Reaction time, movement time, and intelligence. *Intelligence, 3,* 121-126.

Lloyd, F., & Pidgeon, D. A. (1961). An investigation into the effects of coaching on nonverbal test material with European, Asian and African children. *British Journal of Educational Psychology, 31,* 145-151.

Melis, C. J. (1997). *Intelligence: A cognitive-energetic approach.* Academic dissertation. Wageningen: Ponsen & Looijen.

Neisser, U., Boodoo, G., Bouchard, Th. J., Jr, Boykin, A. W., Brody, N., Ceci, S. J., Halpern, D. F., Loehlin, J. C., Perloff, R., Sternberg, R. J., Urbina, S. (1996). Intelligence: Knowns and unknowns. *American Psychologist, 51,*77-101.

Nettlebeck, T. (1982). Inspection time: index for intelligence? *Quarterly Journal of Experimatal Psychology, 34A,* 299-312.

Piaget, J. (1947). *The psychology of intelligence.* London: Routledge & Kegan Paul.

Raven, J. C. (1938). *Progressive Matrices: A perceptual test of intelligence.* London: Lewis.

Reed, T. E., & Jensen, A. R. (1992). Conduction velocity in a brain nerve pathway of normal adults correlates with intelligence level. *Intelligence, 16,* 259-272.

Shepard, L. A., Camilli, G. & Averill, M. (1981). Comparison of procedures for detecting test-item bias with both internal and external ability criteria. *Journal of Educational Statistics, 6,* 317-375.

Sonke, C. J. (2001). Cross-cultural differences on simple cognitive tasks: A psychophysiological investigation. PhD Dissertation. Tilburg University.

Spearman, C. (1927). *The abilities of man.* New York: Macmillan.

Sternberg, (1977). *Intelligence, information processing, and analogical reasoning: The componetial analysis of human abilities.* Hillsdale, NJ.: Erlbaum.

Thurstone, L. L. (1938). Primary mental abilities. *Psychometric Monographs, No. 1.*

Van de Vijver, F. J. R., & Leung, K. (1997a). *Methods and data analysis for cross-cultural research.* Sage Publications, Inc. California, USA.

Van de Vijver, F. J. R., & Leung, K. (1997b). Methods and data analysis of comparative research. In J. W. Berry, Y. H. Poortinga, & J. Pandey (Eds.), *Handbook of cross-cul-*

tural psychology (2nd ed., vol. 1, pp. 257-300). Boston: Allyn & Bacon.

Vernon, P. A. (1969). *Intelligence and cultural environment.* London: Methuen.

Vernon, P. A. (1983). Speed of information processing and general intelligence. *Intelligence, 7,* 53-70.

Vernon, P. A. (Ed.) (1987). *Speed of information processing and general intelligence.* Norwood, NJ: Ablex.

Vygotsky, L. S. (1962). *Thought and language.* Cambridge, MA: MIT Press.

Chapter 1

Cognitive Assessment in Education in a Multicultural Society

Michelle Helms-Lorenz
Fons J.R. van de Vijver

(European Journal of Psychological Assessment, 1995, vol 11, p158-169)

Abstract

The question is raised whether instruments used for cognitive assessment in educational settings such as school achievement tests and intelligence tests are adequate for a multicultural society. Empirical studies often show that migrant pupils score consistently lower on these tests than native pupils. Various factors are discussed that can challenge the equivalence (and hence, the comparability) of the test scores obtained in these groups such as intergroup differences in verbal skills, in cultural values and norms, and in test-wiseness. Commonly applied remedies to enhance the suitability of cognitive tests are discussed: adaptation of existing tests, the use of different norms, statistical and linguistic procedures to correct for item bias, and the development of new tests. Conclusions and implications are discussed.

Key words: COGNITIVE ASSESSMENT, MULTICULTURAL SOCIETY, TEST BIAS, EDUCATION, and ACCULTURATION

Recently several Western societies that were relatively monocultural have become more multicultural; West-European nations are good examples. This transformation brings new challenges in many fields of everyday life. For instance, a steadily increasing number of migrant pupils are entering the schools of these countries each year. The term "migrant" refers here to a broad category of individuals coming from many different parts of the world. This group is not only heterogeneous with respect to their countries of origin, but also with respect to their motives for migration. Some individuals migrate to be reunited with their families who are already living in the host country. Others seek political asylum or flee from war, famine, or political instability. These different causes of migration often imply different expectations for their own future and for their stay in the host country. For some migrants the basis for staying in the safe haven vanishes when the danger diminishes in the home country. Others want to build a new life in the host country and will not return to their original country, at least not in the near future. Finally, migrants are heterogeneous in terms of their knowledge of the dominant language and culture. The multiple heterogeneity of the group creates a major challenge to education in many countries.

The present article focuses on the implications of this change for psychological and educational assessment. Our aim is to scrutinize the role of cognitive tests in multicultural school settings. Most illustrations will refer to Western Europe, in particular the Netherlands, as our primary frame of reference although the issues described are of a more global nature. The first part describes the limited feasibility of most regular cognitive tests for multicultural groups. We argue that the scope of many tests is limited by implicit and explicit references to the dominant language and culture when knowledge of these aspects is not the subject of the test. The second part of this article presents an outline of possible remedies to overcome these limitations. Conclusions and implications are described in the last part.

Cultural Bias and Validity of Inferences from Test Scores

An evaluation of the suitability of an instrument in a multicultural context amounts to an answer to two questions. First, the presence of bias in the instrument should be examined. Three kinds of bias can be envisaged (cf. Van de Vijver & Poortinga, 1995): construct bias, method bias, and item bias (or differential item functioning). Construct bias occurs when the psychological construct measured does not show a complete overlap across cultural groups. For example, everyday conceptualizations of intelligence, notably in non-Western countries, not only include reasoning and knowledge but also social aspects such as the ability to deal with socially complex situations. Whereas the former aspect is usually well represented in Western intelligence tests such as the Raven test, the latter aspect is hardly covered. Method bias refers to the influence of a cultural factor on the test scores such as differential stimulus familiarity that is shared by most or even all items. Whereas method bias refers to anomalies at the test level, item bias refers to problems at the item level that are systematic though unintentional, such as a poor item translation.

Second, suitability in a multicultural context is not an intrinsic property of the test itself but rather depends on the inferences made on the basis of test scores. If the performance on the Raven Test is used to predict scores on a parallel version of the test, bias is less likely than when the Raven test score is used to predict future school success. Broader domains of generalization require more elaborate validation, because each of the three kinds of bias is more likely to occur.

Are Ability, Aptitude, and Achievement Tests Adequate Instruments in Multicultural Societies?

The psychological and educational tests used in education are often divided into achievement, aptitude, and ability tests (e.g., Altink, 1991):

> Aptitudes rely less on specific learning experiences than do achievement tests, but are more related to previous learning experiences than ability measures. ... These tests operationalize skills such as "insight," "understanding" and "problem-solving" with problems from specific subject areas. (p. 253)

Learning potential tests are good examples of aptitude tests (e.g., Hamers, Sijtsma, & Ruijssenaars, 1993). School achievement tests are primarily meant to assess intellectual knowledge and skills acquired in education. This is called crystallized intelligence Cattell and Butcher (1968). Ability tests are supposed to rely the least on previous learning experiences (Drenth, 1979), but research has shown that these tests typically contain elements of both aptitude and achievement tests:

> Traditional intelligence tests, i.e., those that are now most firmly established in the field, and that involve some verbal ability and scholastic knowledge, are mixtures of crystallized and fluid intelligence. (Cattell & Butcher, 1968, p. 20)

Intelligence tests are the best-known example of ability tests. Some intelligence tests are fairly close to aptitude tests; thus, in Raven's tests a deliberate attempt was made to use

simple stimulus material that is not acquired in school. Other intelligence tests, in particular the omnibus intelligence tests such as the WISC-R have subtests in which the presence of knowledge is assumed that can be acquired in school (such as vocabulary subtests).

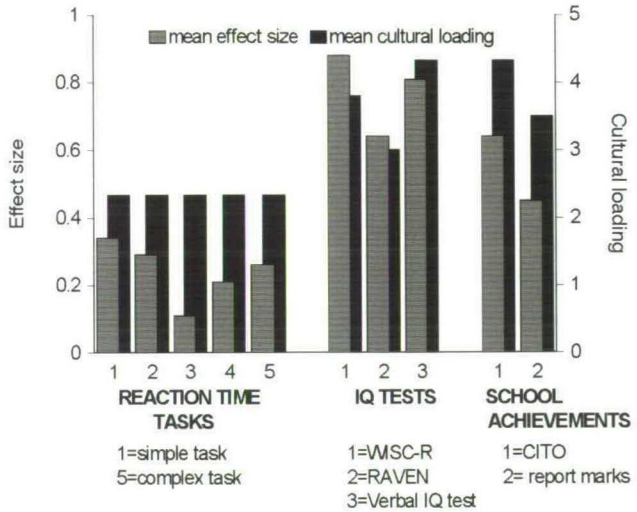


Figure 1. Effect sizes and cultural loadings of cognitive tests

Ability, aptitude, and achievement tests are regularly applied in primary schools in many countries for, among other things, progress testing, assessment of learning difficulties, school advice, and vocational guidance. When migrant and native pupils' mean scores on regular aptitude and ability tests are compared, the latter group usually shows lower mean scores (e.g., De Jong, 1987; Resing, Bleichrodt, & Drenth, 1986). For instance, in the Netherlands the difference in mean score on intelligence tests is usually somewhere between 7 and 15 IQ points. School achievement tests yield similar results (cf. De Jong, 1987; Van Esch, 1983). Figure 1, derived from internal publications by Van de Rijt (1990), Wagenmakers (1994), and Willemse (1989) depicts the effect sizes (i.e., absolute mean score differences divided by the pooled standard deviation) in score comparisons between Dutch natives and migrants for the following types of tests: reaction time tasks (a cognitive test that is described in more detail in a later section), intelligence tests, and school achievement measures. Figure 1 shows that the greatest difference between migrants and natives is found on intelligence tests with a high verbal content. School reports show a smaller difference than would be expected. It has been demonstrated repeatedly (e.g., De Jong, 1987; Resing, Bleichrodt, & Drenth, 1986) that many teachers in the Netherlands judge pupils' performances differentially. Therefore, the psychological meaning of grades probably could differ across the groups of pupils. In the native group grades reflect the child's relative position in the group, whereas the grades of the migrant children are more an indication of their individual progress (Van de Vijver & Willemse, 1991).

Cultural Loadings of Tests

In an evaluation of the adequacy of measurement instruments in multicultural settings the test's cultural loading plays an important role. Cultural loading is a generic term for explicit or implicit references to a specific cultural context, usually the culture of the test composer, in the instrument or its administration. Van de Vijver and Poortinga (1992) distinguish five potential sources of cultural loadings:

-the tester (e.g., when tester and testee are of a different cultural background);
-the testees (e.g., intergroup differences in educational background, scholastic knowledge, and test-wiseness);
-tester-testee interaction (e.g., communication problems);
-response procedures (e.g., differential familiarity with time limits in test procedures);
-cultural loadings in the stimuli (e.g., differential suitability of items for different cultural groups due to stimulus familiarity).

In Figure 1, the mean cultural loadings of the various tests are depicted for the study of Van de Rijt (1990). The mean cultural loadings (evaluated on a 5-point Likert scale) of the various tests were based on ratings by three experts in our department. The correlation between the mean effect size and the mean cultural loading is significant ($r = .73$, $p < .01$). As the cultural loading of the tests increases, the difference in performance between natives and migrants increases.

Cultural loadings have figured prominently in the history of cross-cultural assessment. Thus, the "culture-fair" and the "culture-free" psychometric traditions attempted to reduce or eliminate cultural loadings in tests. This point of view has frequently been criticized by those who believe that stimulus material will always be susceptible to differences in cultural backgrounds of testees (Frijda & Jahoda, 1966).

Van de Vijver and Poortinga (1992) argue that cultural loadings need not be detrimental. The desirability of cultural loadings in measurement procedures is determined by the intention of the test in question: "the (un) desirability in assessment instruments depends on the generalizations (inferences) envisaged on the basis of the scores" (p. 22). The elimination of biased items does not necessary imply an increase in the predictive validity of the test. The elimination of such items could just as well mean that items that correlates with future school achievement for migrant pupils are removed because future school achievement itself has a high cultural loading. If a particular test is intended to test knowledge gained during a course at school (achievement test), it is quite likely that culture-specific knowledge is tested (e.g., history, geography). In this case, cultural loading is unavoidable and even desirable. In a multicultural class, intergroup differences reflect differential mastery of the subject matter. More generally, a distinction can be made between generalizations about achievements (past, present, and future behaviors) and about abilities and aptitudes. In the latter case, cultural loadings are usually undesirable. Inferences about intergroup differences in aptitudes and abilities that are based on common Western instruments can have a dubious validity. Such

tests may suffer too many shortcomings to enable cross-cultural comparisons. These will be discussed in the next paragraph.

Validity Threats of Aptitude and Ability Tests in Multicultural Settings

Subject-related factors that can differentially influence test performance of natives and migrants are: verbal abilities, cultural norms and values, test-wiseness, and acculturation strategy. These factors can cause bias and reduce the validity of inferences drawn from score comparisons.

Conventional mental tests often call for high *verbal abilities and skills*. Migrant pupils usually differ from native pupils in native language and cultural knowledge and skills. Test instructions as well as the item phrasings can contain words or specific idioms that unintentionally discriminate between natives and migrants. The use of idioms requires special attention because an idiom's meaning will often be clear to natives but unclear to migrants. A literal translation will not convey the meaning of an idiomatic expression, and such expressions are often mastered fairly late in the acquisition of a second language. The problem is particularly salient when verbal ability itself is not the subject of the test; for example, in embedded arithmetic exercises word knowledge can easily become an unintended source of score differences between natives and migrants.

Cultural norms and values can also be introduced unintentionally into tests; to respond correctly to such items requires a great awareness of the dominant culture. The following item of the WISC-R illustrates this point: "What is bacon?" It has been demonstrated that Turkish and Moroccan pupils have difficulty with this item (Van de Rijt, 1990). This is not so surprising considering the fact that these pupils are brought up in an Islamic culture where eating pork is taboo. "What is bacon?" not only measures vocabulary but is also a measure of acquaintance with native customs. The problem that tests measure the degree of assimilation to the native way of life is not restricted to the WISC-R. Another example is a Binet item that asks the child to pick the prettier of two faces. Critics complain that the judgment is "loaded with white middle class values" (Jensen, 1980, p. 5). In the previous examples references to the dominant cultures can be easily discerned. In many cases, however, the references are more subtle and difficult to spot. A committee of experts has scrutinized the most common psychological tests in the Netherlands; they concluded that all tests contain, often implicit, references to Dutch norms and values (Hofstee, 1990; Hofstee, Campbell, Eppink, Evers, Joe, Van de Koppel, Zweers, Choenni, & Van de Zwan, 1990).

Test-wiseness (Sarnacki, 1979) can differentially influence test performance. A well-known example is that the ability to perform well on multiple-choice tests often presupposes good linguistic skills necessary to understand the differences between the response alternatives. Another example is the ability to deal with time limits in speed tests. The major problem is finding an optimal combination of speed and accuracy. Prior experience with working under time constraints can help a child master this skill.

Native pupils as well as migrant pupils with a substantial educational history in the same culture will often have similar and extensive experience in dealing with psychological and educational tests. However, first generation pupils from different educational backgrounds may not have mastered these skills.

Finally, *acculturation strategy and expectations about one's future* can have a bearing on the performance of migrant pupils. When individuals migrate to a new country, acculturative stress is evoked. Adaptation to the new situation can come about in different ways. The adaptation styles are referred to as acculturation strategies in the literature. Four styles are commonly distinguished: "assimilation," "integration," "separation," and "marginalization" (e.g., Berry, 1994). The different styles are characterized by distinctive attitudes towards their own culture as well as the other culture, commonly the majority group of the host country.

Persons who value relationships with individuals of the new culture, and who also regard the relationship with their own culture as nonessential, assimilate rapidly and experience little or no stress. Integration is the acculturation strategy whereby maintaining one's own culture and simultaneously developing positive relationships with members of the dominant culture are regarded as important. Integration is associated with a bicultural identity. The acculturation strategy in which a person has no intention of having positive relationships with members of the new culture and who values their own culture and relationships with its members is called separation. Finally, the most stressful acculturation strategy is marginalization. This style occurs when an individual does not wish to have relationships with members of either culture, i.e., both cultures are rejected. According to Boski (1994), Berry's approach is too general and hardly pays attention to specific similarities and differences of the native and host culture. Berry's model barely touches on cultural distance, an important variable in acculturation processes.

Berry and Boski have delineated different mediating factors underlying the acculturation process. Berry has identified the following factors: acculturation strategy, expectations, prior knowledge of the language and culture of the dominant group, migration motivation (push vs. pull), life changing events perceived as opportunities or as problems, initial health, age, and education, ability to communicate with the other culture, coping strategies and resources, perceived stressors, status, appraisal/reaction to societal attitudes and use made of social support. Boski suggests the following mediating factors: cultural distance, time spent in the host country, relationship to one's country of birth and of primary socialization. Approval of the home country's cultural values is detrimental to adaptation to the host country.

In our view, acculturation strategies play an important role in educational settings. A person's acculturation strategy is related to his/her expectations for the future. A migrant who is planning to emigrate from the host country in the near or distant future will most probably be less willing to invest great effort in learning the local language and customs. This person can quickly reach a level of knowledge and skills that will

meet daily needs. If the immigration is more or less permanent, the payoffs for learning the language and customs will be greater. A synthesis of the acculturation models such as those proposed by Berry and Boski might bring us a step closer to a full-fledged theory of psychological acculturation.

From a theoretical point of view, subject-related factors can lead to all three forms of bias described above. Yet, not all three sources of bias are equally likely. Construct bias is far less likely in school achievement tests than in aptitude and ability tests. The most probable kind is method bias, because most subject-related factors such as intergroup differences in verbal skills and test-wiseness will affect all items in a more or less uniform way, thereby introducing invalid intergroup differences in average test performance.

Increasing the Suitability of Tests for Multicultural Settings

A number of procedures are available to reduce or even eliminate problems encountered when measuring cognitive abilities in multicultural settings. Below we discuss the adaptation of existing tests, the application of different norms, statistical and linguistic procedures, and the development of new tests (cf. Van de Vijver, Willemse, & Van de Rijt, 1993).

Adapting existing tests

First, existing tests can be adapted to enhance their suitability in a multicultural context (Hambleton, 1994; Schwarz, 1961). The rationale behind this approach is that tests developed and standardized for one specific cultural group are not necessarily adequate for another cultural group due to the explicit and implicit references to the cultural background of the test composer. Various adaptations have been proposed in the literature such as giving clear and lengthy instructions, providing exercises after the examples, and avoiding complicated grammatical structures and local idioms (cf. Van de Vijver & Leung, 1995). There are numerous examples of test adaptations in the literature. Bravo, Woodbury-Farina, Canino, and Rubio-Stipec (1993) developed a Spanish translation and adaptation of the Diagnostic Interview Schedule for Children. The adaptation was aimed at identifying phenomena similar to those of the original English version in another cultural context. To attain cross-cultural equivalence, five dimensions were addressed: semantic, technical, content, criterion, and conceptual. The translation and adaptation process involved various methodological steps including a translation by a bilingual committee, back-translation, and reliability and validity testing. Reliability and validity were assessed using a sample of 248 children (aged 9-17 yrs) drawn from the community and a sample of 74 children selected from special treatment populations in Puerto Rico. Results suggest that the adapted instrument is measuring phenomena related to dysfunction in social, psychological, and academic realms.

Most test adaptations reported in the literature are aimed at enhancing the validity of an instrument for a particular group. There are only a few examples in which test

adaptations are implemented to enhance the appropriateness of the test in a multicultural setting. The latter kind of test adaptation is more involved than the former. The work by Resing, Bleichrodt, and Drenth (1986) is an example of the latter method. They studied the suitability of the Revised Amsterdamse Kinder Intelligentie Test (RAKIT), an intelligence test for migrant children, that had been standardized previously for the native Dutch population.

Differential Norms

This remedy entails different interpretations of the same scores for different cultural groups; for example dealing with various cutoff scores in job application procedures. Differential norms are often used to compensate for social inequality and unequal opportunities. There are several ways in which differential norms can be applied. Thus, it is possible to choose different pass-fail cutoff points for different cultural groups, or to designate beforehand a fixed percentage of migrants to progress to higher educational levels without considering the average level of this group. The application of group-dependent norms is often part of social or political programs such as positive discrimination, equal opportunities, and affirmative action.

Sackett and Wilk (1994) discuss three rationales for score adjustment: to attain business or social goals, to alleviate test bias, and to obtain fairness. The first justification is based solely on social concerns and is independent of technical merits of the instrument in question. The second position is a technical (statistical) issue. The authors argue that score adjustment should be permitted when bias is detected. The third justification focuses on a fair selection system rather than the individual test. A selection system is deemed unfair if the minority selection rate is less than the rate that would be obtained if selection were based on actual job performance. The authors summarize research findings on cognitive ability tests for personnel selection and conclude that these instruments show consistent predictive validity for a wide range of jobs, a lack of predictive bias against Blacks and Hispanics, and large, consistent adverse impact by race.

An example of this approach has been developed for the WISC-R by Mercer (1979). She developed a System for Multicultural Pluralistic Assessment. Her procedure was criticized by Cronbach (1984, pp. 209-214) for various methodological reasons such as small sample sizes, lack of geographical representativeness, and most importantly, lack of empirical evidence that the IQs derived from her way of scoring the WISC-R has a higher predictive validity than the common scoring method.

The use of differential norms is an area in which science and social policy meet. Views on the acceptance of the application of differential norms cannot be taken for granted. Sackett and Wilk (1994) and Gottfredson (1994) discuss social and political perspectives, as well as scientific and theoretical issues, concerning various methods for score adjustment in the U.S. Moreover, substantial cross-cultural differences in the acceptance of positive discrimination by the general public (i.e., the dominant cultural

group) are likely to be found. In the Netherlands the public opinion seems to be more favorable toward the application of differential norms in educational settings than in the labor market. At the end of primary school an achievement test (CITO Eindtoets) is administered. Test scores combined with the teacher's judgment of the student's capacities, form the basis of a recommendation for the most suitable type of secondary school for the child. There is empirical evidence that when the test performance of natives and migrants is equal, the latter tend to be advised to seek an intellectually more demanding type of school (De Jong, 1987). Such a differential treatment is less likely to be accepted on the labor market in the Netherlands.

Opinions held by scientists and social policy makers regarding fair test use can differ markedly. An example can be found in the Golden Rule Settlement. This example illustrates how selection procedures can be driven by the public's opinion of fairness. The settlement between the Golden Rule Insurance Company and the Illinois Department of Insurance and Educational Testing Service concerns a system for determining which items would be included in the Illinois insurance licensing examination. A raw difference of .15 or more in an item's p-value, favoring White applicants over Black applicants, was the criterion used to identify items that should not be included in the test. Holland and Wainer (p. 15, 1993) present two lines of evidence to support the psychometric view that a p-value difference by itself is not a sufficient reason for concluding that an item is biased. They argue that large differences in p-values are expected given the historical differences in education (i.e., nature, quality, and length of schooling) between Blacks and Whites; furthermore, the removal of a legitimate part of the test would lower its validity (cf. Faggen, 1987). Holland and Wainer question the legitimacy of the underlying psychometric procedure of (many) item bias techniques that match groups according to ability levels. This matching criterion produces a group of unrepresentative Blacks and a group of unrepresentative Whites to be compared.

Another dilemma between science and social policy is presented in the following example. In the U.S. the issue of subgroup norming has been controversial for more than a decade. Until recently, the General Aptitude Test Battery (GATB) was frequently used to screen applicants for many jobs in the USA. Each registrant's score was calculated as a percentile score within his/her racial or ethnic group. According to Reynold (Assistant Attorney General of Civil Rights in the U.S. Department of Justice, 1986), this practice constituted an illegal and unconstitutional violation of an applicant's right to be free from racial discrimination. The controversy reached a new peak with the passage of the Civil Rights Act of 1991, which banned any form of "score adjustment" on the basis of "race, color, religion, sex, or national origin" (Pub. L. No. 102-166, Section 106). Others contend that this law is insufficiently backed by scientific evidence (Gottfredson, 1994) and could legitimize discrimination in selection procedures, which is prohibited by yet another law. Judicial constraints could have the unintended and undesirable consequence that only those assessment procedures that are unstructured (such as the job interview) will be applied, merely because it is difficult or virtually impossible to show their (un)

fairness. However, such procedures tend to have a low validity. Thus, a one-sidedness approach to banning discrimination, however desirable from the viewpoint of social policy, may indirectly have adverse effects on the validity of the selection procedure.

Statistical and Linguistic Procedures

A third possibility entails statistical and linguistic procedures to improve the suitability of instruments in a multicultural setting. This tradition is known as "item bias" (e.g., Berk, 1982) and "differential item functioning" (e.g., Holland & Wainer, 1993). This approach is more specific than the two mentioned in the previous paragraphs. Whereas test adaptations and the use of differential norms concern the test as a whole, "item bias" focuses on the usefulness of test items. After the test is administered to members of different cultural groups, each item is scrutinized. This can be accomplished with linguistic (De Jong & Vallen, 1989) or with psychometric (Holland & Wainer, 1993; Kok, 1988) procedures. A recent example of linguistic analysis can be found in a study conducted by the Dutch Test Screening Committee mentioned above (Hofstee, 1990; Hofstee, Campbell, Eppink, Evers, Joe, Van de Koppel, Zweers, Choenni, & Van de Zwan, 1990). As another example, the Fawcett Society (1987) examined a range of exam papers and identified several types of (sex) discrimination in the item formulations.

Psychometric analysis of item bias has proliferated in the last few decades. A wide range of techniques has been developed; a review can be found in Berk (1982) and Holland and Wainer (1993). Item bias is believed to exist when persons from different cultural groups with the same ability level have an unequal probability of responding correctly to an item. A schematic outline of statistical techniques used to study score equivalence (the absence of bias) can be found in Van de Vijver and Poortinga (1991). Most item bias studies have been conducted in the U.S.; the number of studies carried out in Western Europe is very limited.

Developing New Instruments

Finally developing new instruments can circumvent shortcomings of common tests. The idea of designing instruments that may be used in cross-cultural comparisons is not new. Examples of initiatives to develop new instruments can be found in the culture-free and culture-fair test movements. Cattell's Culture Fair Intelligence Test (Cattell and Cattell, 1963) and Raven's Progressive Matrices (Raven, 1938) are products of this approach. More recent endeavors to develop cross-culturally equivalent adequate paper-and-pencil instruments have focused primarily on aptitude measurement. For example, in the last decade in the Netherlands new learning potential tests have been developed and validated (e.g., Hessels & Hamers, 1993). Learning potential tests are based on the principle of measuring "learning potential by giving the child a standard task and observe how fast he learns it" (Jensen, 1961, p. 148). The "Leertest voor Etnische Minderheden" (LEM; Hessels & Hamers, 1993) is based on Vygotsky's theory of "the zone of proximal development" and consists of subtests for classification, word-

object association, number series, syllable recall, and figurative analogies. Subtest scores are based on the extent to which children benefit from help (the more help needed the lower the score). The effect of culture was reduced by eliminating inappropriate test content, minimizing the influence of test-wiseness by using familiarization and training, providing appropriate samples for local norms, and reducing language bias by using non-verbal instructions. The migrant children's performances on the LEM differed significantly from that of native children but this difference was smaller than with traditional IQ-scores. Furthermore, the LEM was found to discriminate well in the low ability range, which implies that the LEM may prevent children from being incorrectly labeled as mentally retarded.

In the following section a detailed description is given of a new test, developed at the Tilburg University and serves to illustrate its advantages. In our research we do not use traditional paper-and-pencil tests but rather administer tasks that are similar to the so-called elementary cognitive tasks (e.g., Vernon, 1987). The focus is on speed rather than on accuracy; test items are so simple that all subjects can answer them correctly. In developing this instrument the most important objective is to reduce the influence of potentially biasing *subject*-related factors on the test performance, such as verbal skills, norms and values, and test-wiseness of the testees. The test is virtually nonverbal. The interaction between tester and testee is reduced as much as possible, and the role of the tester in test administration is marginal compared to his/her role in the administration of traditional paper-and-pencil tests. Furthermore, we have attempted to reduce *instrument*-related biasing factors such as the cultural loadings of test items. The stimulus material consists of simple geometric figures. In order to become familiar with the stimulus material and the test setting, subjects receive ample opportunity for practice.

The tester starts by giving instructions (in simple words or by pantomime) and by demonstrating a few items. The testee sits in front of a computer monitor. In the first version of the test, the testee responded by pressing a response button device (see Figure 2); in a more recent version of the battery a mouse is used.

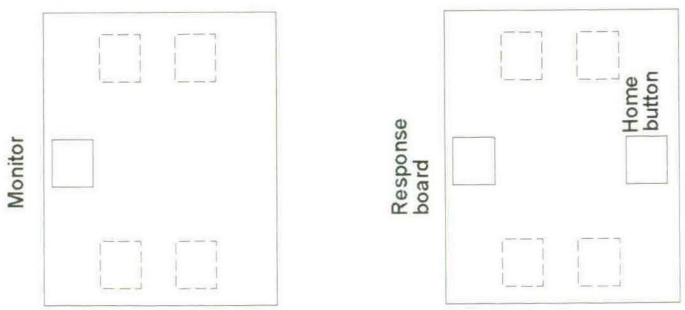


Figure 2. Experimental setup for tasks 2-5 of the reaction time test

Willemse (1989; Van de Vijver & Willemse, 1991) administered the computerized test battery to native and migrant pupils. The battery consists of five tasks of increasing cognitive complexity. The first task is a simple reaction time task. Two response buttons are visible (see nondashed buttons in Figure 2), a home button and a response button. After an auditory warning signal, the outline of a square appears on the screen. At this point the subject is instructed to press the home button. A few seconds later the square on the screen becomes black. The subject is asked to push the top button as soon as this change occurs. This task (as well as the others) consists of 20 trials. Four additional tasks are choice reaction time tasks, which require the use of all response buttons. In the second task, five squares appear on the screen (cf. Figure 2). After a few seconds, one of the squares becomes black. As soon as this happens the testee is supposed to press the corresponding response button. The third task consists of five squares, four of which are identical (for example, one of the geometric patterns shown in Figure 3 appears on four squares). The fifth square consists of a different geometric pattern. The testee is required to press the response button of the unique geometric pattern. In the fourth task, two pairs of identical figures appear on the monitor and a different pattern appears on the fifth square. The testee's task is to find the "odd-one-out" and press the corresponding response button.

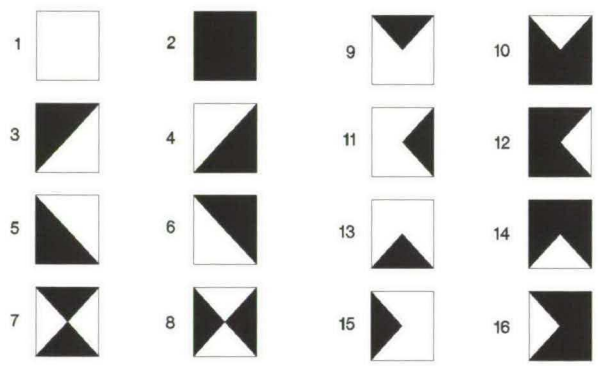


Figure 3. Figures used as stimulus material

The last task introduces "complementary" squares. Two figures are complementary if they form a full square when joined. Two pairs of complementary squares are presented in each row of Figure 3. Two pairs of complementary squares appear accompanied by one figure with no complement. The pupil is asked to press the button corresponding to the non-complementary square as fast as possible.

The results of research carried out by Willemse (1989), later replicated by Van de Rijt (1990), showed that both native and migrant pupils had the same performance level for all tasks. Furthermore, as the tasks increased in complexity, the correlation

with school achievement became stronger. The simplest tasks yielded no significant relationship with school achievement; correlations for the most complex tasks were around -.50 in both studies (the correlation is negative, indicating that pupils with faster reaction times have higher school achievements).

These new test procedures may provide a viable alternative to conventional tests in various situations. Conventional tests are impractical when the pupil has an insufficient mastery of the local language. This may occur when the pupil has recently immigrated or when the language spoken by the child at school and at home is not identical. Educational guidance may provide another area of application. If school progress is slow, the instruments described above may provide insight into the role of intellectual factors in educational problems.

Conclusion

The shift from mono- to multicultural school populations in many Western countries has raised the question of the adequacy of conventional school achievement, aptitude, and ability tests. Research has indicated that migrants consistently score lower on these tests than natives do. The differences are however, not the same for all types of tests. In our work we found that the largest differences occur in tests with high cultural and verbal loadings, for example in traditional intelligence tests.

We have argued that the question of suitability primarily depends on the possible presence or absence of construct, instrument, and item bias and secondly, on the intended purpose of the test scores. Construct, instrument, and item bias could occur in all the types of tests mentioned but are not equally likely for each type. Construct bias is far less likely to occur in school achievement tests than in aptitude and ability tests. The most probable bias in all types of tests is method bias because most subject-related factors such as intergroup differences in verbal skills and test-wiseness will affect all items in a more or less uniform way, thereby inducing intergroup differences in average test performance that cannot be attributed to the construct of the test.

Testing pupils can have several goals. If the intention is to predict future school performances, culturally loaded items may well be useful because these items contain the same cultural loading of the school and circumstances in which the child is to perform. If the purpose is to generalize about abilities or aptitudes, cultural loadings are undesirable.

The remedies discussed are designed to solve different types of bias. If a construct does not show a complete overlap across cultural groups, the test could be restricted to the common aspects; culture-specific aspects can be excluded or separately assessed. The procedure is adequate as long as the domain restriction is acknowledged. To our knowledge this remedy is not frequently applied. The influence of method bias is underrated and insufficiently studied. The administration of instruments ostensibly measuring the same construct with varying degrees of cultural entrenchment provides insight into the presence or absence of method bias.

Statistical and linguistic bias techniques can identify item bias. It is regrettable that these techniques are infrequently applied. The use of different norms for cultural groups is politically and socially delicate. The development of new ability tests could reduce or even eliminate some of the problems encountered with conventional tests. More care should be taken in the operationalization of the construct to be measured. Furthermore, instrument and subject factors that may threaten test validity in a multicultural setting should be identified and minimized.

It would be naive to assume that all problems encountered with the use of psychological tests for migrant pupils can be solved with these remedies. Furthermore, new assessment procedures such as the reaction time tests described above do not render conventional test superfluous. Both types of tests appear to have different applications. Conventional tests may be better predictors of future school success, whereas innovative procedures may provide better insight into the intellectual capacities of migrant pupils. These are both important goals in educational settings. Finally, noncognitive factors such as acculturation styles influence the school performance of migrant pupils. The assessment of migrant students' cognitive abilities should take into account the individuals' future expectations. It is only through a balanced treatment of all issues involved that psychology can meet the challenge of multiculturalism in education.

References

Altink, W. M. M. (1991). Admission for preentry science upgrading courses in Southern Africa. Choice of selection instruments. *Journal of Cross-Cultural Psychology, 22*, 250-272.

Berk, R. A. (Ed.) (1982). *Handbook of methods for detecting item bias*. Baltimore: Johns Hopkins University Press.

Berry, J. W. (1994). Acculturation and psychological adaptation: An overview. In A. Bouvy, F. J. R. Van de Vijver, P. Boski, & P. Schmitz (Eds.), *Journeys into cross-cultural psychology* (pp. 129-141). Lisse: Swets & Zeitlinger.

Boski, P. (1994). Psychological acculturation via identity dynamics: Consequences for subjective well-being. In A. Bouvy, F. J. R. Van de Vijver, P. Boski, & P. Schmitz (Eds.), *Journeys into cross-cultural psychology* (pp. 197-215). Lisse: Swets & Zeitlinger.

Bravo, M., Woodbury-Farina, M., Canino, G. J., Rubio-Stipec, M. (1993). The Spanish translation and cultural adaptation of the Diagnostic Interview Schedule for Children (DISC) in Puerto Rico. *Journal of Culture Medicine and Psychiatry, 17*, 329-344.

Cattell, R. B., & Butcher H. J. (1968). *The prediction of achievement and creativity*. Bobbs-Merrill Company, New York.

Cattell, R. B., & Cattell, A. K. S. (1963). *Culture fair intelligence test*. Champaign, IL: Institute for Personality and Ability Testing.

Civil Rights Act of 1991, Pub. L. No. 102-166, 105 Stat. 1071 (Nov. 21, 1991).

Cronbach, L. J. (1984). *Essentials in psychological testing* (4th ed.). New York: Harper & Row.

De Jong, M. J. (1987). *Herkomst en kansen. Allochtone en autochtone leerlingen tijdens de overgang van basis naar voortgezet onderwijs*. Lisse: Swets & Zeitlinger.

De Jong, M., & Vallen, T. (1989). Linguïstische en culturele bronnen van itembias in de Eindtoets Basisonderwijs van leerlingen uit etnische minderheidsgroepen. *Pedagogische Studiën, 66,* 390-402.

Drenth, P. J. D. (1979). Prediction of school performance in developing countries: School grades or psychological tests? *Journal of Cross-Cultural Psychology, 8,* 49-70.

Faggen, J. (1987). Golden Rule revisited: Introduction. *Educational Measurement, Issues and Practice, 6* (Summer), 5-8.

Fawcett Society (1987). *Exams for the boys.* Hemel Hempstead: The Fawcett Society.

Frijda, N., & Jahoda, G. (1966). On the scope and methods of cross-cultural research. *International Journal of Psychology, 1,* 109-127.

Gottfredson, L. S. (1994). The science and politics of race-norming. *American Psychologist, 49,* 955-963.

Hambleton, R. K. (1994). Guidelines for adapting educational and psychological tests: A progress report. *European Journal of Psychological Assessment, 10,* 229-244.

Hamers, J. H. M., Sijtsma, K., Ruijssenaars, A. J. J. M. (Eds.) (1993). *Learning potential assessment. Theoretical, methodological and practical issues.* Lisse: Swets & Zeitlinger.

Hessels, M.G.J., & Hamers, J.H.M. (1993). A learning potential test for ethnic minorities. In J. H. M. Hamers, K. Sijtsma, & A. J. J. M. Ruijssenaars (Eds.), *Learning potential assessment. Theoretical, methodological and practical issues.* Lisse: Swets & Zeitlinger. 285-313.

Hofstee, W. K. B. (1990). Toepasbaarheid van psychologische tests bij allochtonen. *De Psycholoog, 25,* 291-294.

Hofstee, W. K. B., Campbell, W. H., Eppink, A., Evers, A., Joe, R. C., Koppel, J. M. H. van de, Zweers, H., Choenni, C. E. S., Zwan, T. J. van de (1990). *Toepasbaarheid van psychologische tests bij allochtonen.* LBR reeks, nr. 11. Utrecht: LBR.

Holland, P. W., & Wainer, H. (Eds.) (1993). *Differential item functioning.* Hillsdale, NJ: Erlbaum.

Jensen, A. R. (1961). Learning Abilities in Mexican-American and Anglo-American Children. *California Journal of Educational Research, 12,* 147-159.

Jensen, A. R. (1980). *Bias in mental testing.* New York: Free Press.

Kok, F. G. (1988). *Vraagpartijdigheid. Methodologische verkenningen.* Amsterdam: University of Amsterdam.

Mercer, J. R. (1979). *System of multicultural pluralistic assessment (SOMPA): Technical Manual.* New York: The Psychological Corporation.

Raven, J. C. (1938). *Progressive Matrices: A perceptual test of intelligence.* London: Lewis.

Resing, W. C. M., Bleichrodt, N., & Drenth, P. J. D. (1986). Het gebruik van de RAKIT bij allochtoon etnische groepen. *Nederlands Tijdschrift voor de Psychologie, 41,* 179-188.

Sackett, P. R., & Wilk, S. L. (1994). Within-group norming and other forms of score adjustment in preemployment testing. *American Psychologist, 49,* 11, 929-954.

Sarnacki, R. E. (1979). An examination of test-wiseness in the cognitive test domain. *Review of Educational Research, 49,* 252-279.

Schwarz, P. A. (1961). *Aptitude tests for use in developing nations.* Pittsburg: American Institute for Research.

Van de Rijt, B. (1990). *Reactiesnelheidstest. Een aanvulling voor allochtonen op de bestaande intelligentietests.* Unpublished master's thesis, Tilburg University, Tilburg.

Van de Vijver, F. J. R., & Leung, K. (1995). *Methods and data analysis of cross-cultural research.* Handbook of Cross-Cultural Psychology. Manuscript submitted for publication.

Van de Vijver, F. J. R., & Poortinga, Y. H. (1991). Testing across cultures. In R. K. Hambleton & J. N. Zaal (Eds.), *Advances in educational and psychological testing* (pp. 277-307). Dordrecht: Kluwer.

Van de Vijver, F. J. R., & Poortinga, Y. H. (1992). Testing in culturally heterogeneous populations: When are cultural loadings undesirable? *European Journal of Psychological Assessment, 8,* 17-24.

Van de Vijver, F. J. R., & Poortinga, Y. H. (1995). *Towards an integrated analysis of bias in cross-cultural assessment.* Manuscript submitted for publication.

Van de Vijver, F. J. R., & Willemse, G. R. W. M. (1991). Are reaction time tasks better suited for ethnic minorities than paper and pencil tests? In N. Bleichrodt & P. J. D. Drenth (Eds.), *Contemporary issues in cross-cultural psychology.* Lisse: Swets & Zeitlinger. 450-464.

Van de Vijver, F. J. R., Willemse, G. R. W. M., & Van de Rijt, B. (1993). Het testen van cognitieve vaardigheden van allochtone leerlingen. *De Psycholoog, 28,* 152-159.

Van Esch, W. (1983). *Toetsprestaties en doorstroomadviezen van allochtone leerlingen in de zesde klas van lagere scholen.* Nijmegen: Instituut voor Toegepaste Sociologie.

Vernon, P. A. (Ed.) (1987). *Speed of information processing and intelligence.* Norwood, NJ: Ablex.

Wagenmakers, F. (1994). *Reactiesnelheid en intelligentie. Een kwantitatief onderzoek naar de relatie tussen reactietijdtaken en intelligentie bij allochtonen en autochtonen.* Unpublished master's thesis, Tilburg University, Tilburg.

Willemse, G. (1989). *Keuze-reactietijdtaken in multi-culturele context.* Unpublished master's thesis, Tilburg University, Tilburg.

Chapter 2

Cross-Cultural Differences in Cognitive Performance and Spearman's Hypothesis: G or C?

Michelle Helms-Lorenz
Fons J.R. van de Vijver
Ype H. Poortinga

Submitted

Abstract

Common tests of Spearman's hypothesis, according to which performance differences between cultural groups on cognitive tests increase with their g loadings, confound cognitive complexity and verbal—cultural aspects. The present study attempts to disentangle these components. Two intelligence tests and a computer-assisted elementary cognitive task were administered to 474 second-generation migrant and 747 majority-group pupils in The Netherlands, with ages ranging from 6 to 12 years. Theoretical complexity measures were derived from Carroll's (1993) model of cognitive abilities and Fischer's (1980) skill theory. Cultural loadings of all subtests were rated by 25 third-year psychology students. Verbal loading was operationalized as the number of words in a subtest. A factor analysis of the tests' loadings on the first principal component, theoretical complexity measures, and ratings of cultural loading revealed two virtually unrelated factors, called g and c (for culture). The findings suggest that performance differences between majority-group members and migrant pupils are better predicted by c than by g.

Introduction

Spearman (1927) was the first to observe that on tests with a higher g saturation tended to reveal larger performance differences between ethnic groups (p. 379). The g saturation of a test refers to its cognitive complexity.

Elaborating on these observations, Jensen (1985) formulated "Spearman's Hypothesis" (SH), which predicts larger performance differences between ethnic groups on tests with a higher g loading. Performance differences are measured by effect sizes, such as Cohen's d. A test's g loading is usually represented by its loading on the first principal component of the intertest correlation matrix, or by its loading on the first principal component of the second order g factor derived from hierarchical factor analysis (i.e., the general factor among the obliquely rotated first-order factors). A less common measure of g is the use of correlations with tests that have a high g loading. For example, Jensen (1993) has used Raven's Standard Progressive Matrices to calibrate tests of unknown g loadings.

In the discussion of studies on SH a distinction can be made between studies that (1) directly test SH, (2) propose and test alternative explanations of SH, (3) refute alternative explanations of SH, and (4) test the generalizability of SH.

Direct Hypothesis Tests

Jensen (1998) gives an up-to-date account of research into SH based on paper-and-pencil tests and reaction time (RT) tests, most frequently employing samples of African-Americans (AA) and European-Americans (EA). A direct SH test was carried out by Jensen (1985). He selected 11 studies in which batteries of minimally six diverse, reliable tests had been administered to large African-American and European-American samples. One principal component was extracted (which is interpreted as the

g factor), separately for the ethnic groups. The factorial similarity of the group factor loadings was measured by Tucker's phi. Comparisons of factor structures of the two groups tend to yield high values of Tucker's phi (cf. Van de Vijver, 1997, 1999). Differences in mean test scores between the groups were expressed in standard score units (effect size). After disattenuation, differences between the ethnic groups correlated significantly positive with the g loadings of the test (Spearman's $r_S = +.59$, $p < .001$).

Naglieri and Jensen (1987) matched African-American and European-American fourth and fifth graders on age, gender, school, and socioeconomic status. They found a Spearman's rank order correlation of .75 between the test's g loading and the standardized mean differences between the ethnic groups on a diverse set of 24 mental tests.

New studies made a refinement of the SH necessary. In its original formulation, SH implied that EA—AA differences in test scores were exclusively associated with the test's g loading. It was found, however, that when the g factor was removed, significant differences were not completely eliminated; the correlation between [EA—AA] differences and g did not approach unity, even after correction for attenuation (Jensen & Reynolds, 1982). Moreover, short-term memory tasks commonly showed smaller and visualization tasks larger EA—AA differences than would be expected on the basis of SH (Jensen, 1998). Therefore, Jensen (1985) formulated a weaker version of SH, stating that the variation in performance differences of African-Americans and European-Americans on various tests is "associated *predominantly* (rather than exclusively) with the test's g loading" (p. 231, emphasis in original).

Elementary cognitive tasks have also been employed to test SH. Significantly larger EA—AA differences were found on more complex choice RT tests (mean response latency was used as index of complexity) than on simple RT (Jensen, 1982). Higher correlations with IQ and g were found for more complex RT tasks than for simple tasks (e.g., Vernon & Jensen, 1984). However, complexity was operationalized as response latency, and this is questionable; processes that take longer are not necessarily more complex.

In another study, Jensen (1993) administered three elementary RT tasks (simple, choice, and discrimination tasks) to 585 European-American and 235 African-American elementary school pupils of both sexes (mean age 10.9 years). The complexity of the RT tasks was operationalized as their correlation with the score on Raven's Standard Progressive Matrices. This complexity measure was correlated with EA—AA differences per group and for the combined group. The Spearman rank correlation between the complexity measure and performance differences was significantly positive for both groups. After the linear component of Raven's Standard Progressive Matrices variance was regressed out of the elementary cognitive tasks in each racial group, residual scores revealed a remarkable pattern: African-Americans showed faster RTs, meaning that by removing g, the sign of the EA—AA difference is reversed. The authors concluded that g was a suppressor of pure motor aspects of RTs.

A Math Verification Test was administered to a sub sample of 73 European-American and 118 African-American children (Jensen, 1993). Participants had to indicate

whether single-digit additions, subtractions, and multiplications were correct or incorrect; RTs were registered. Correlations between the g loadings (i.e., correlations with Raven's Standard Progressive Matrices) and the EA—AA differences strongly supported SH.

Another test of SH is based on Stankov's (1983) finding that the active (processing) component of working memory is more highly correlated with g than is the passive (storage) component. Jensen (1984) experimentally manipulated g loadings by presenting tasks both separately and simultaneously. He found that, compared to the separate presentation, the simultaneous presentation showed larger EA—AA differences and higher g loadings, thereby supporting SH.

SH was also confirmed in a study in which the Stanford-Binet Intelligence Scale for Children was applied to 66 African-American and European-American children of 3 years of age (Peoples, Fagan, & Drotar, 1995). An analysis of test performance with race as independent variable yielded an F ratio that was used as a measure of performance differences. The correlation between the ratio and g was .71.

Studies Supporting Alternative Explanations of SH

Humphreys (1985) analyzed data of more than 100,000 individuals of the Project Talent Data Bank on a large set of cognitive tests. Data were analyzed for participants of low and high socioeconomic status and for African-Americans and European-Americans separately. Loadings of g correlated .17 with race and .86 with socioeconomic status differences. The performance differences were attributed to adverse environmental factors (low SES) that affect all individuals to the same extent, irrespective of race.

The role of cultural bias has also been explored. A test is said to be culturally biased when test score differences between groups obtained with the instrument are influenced by measurement artifacts. Evidence for the role of cultural bias in the explanation of EA—AA differences comes from Montie and Fagan (1988). In addition to large mean differences favoring European-American preschool children (three-year olds) tested with the third revision of the Stanford-Binet test, these authors found that performances were larger on some items relative to others (significant race x item interactions). They concluded that test bias might have contributed to the racial differences in IQ.

Studies Refuting Alternative Explanations of SH

Jensen (1993) refuted motivational effects as an alternative explanation of EA—AA differences. African-Americans showed faster Movement Times (MT) and slower RTs than European-Americans in elementary cognitive tasks. According to Jensen, it is difficult to see why European-Americans would be more motivated in RT-related processes and less motivated in MT-related processes as both refer to processes that immediately follow each other in the tasks studied and together do not take more than a few seconds.

Strategy differences between African-Americans and European-Americans can also be envisaged as an explanation of SH. Jensen (1993) addressed this question by examining RT:MT ratios. If the two groups show strategy differences, this should be

expressed in different ratios (e.g., depending on the strategy used, the decision about which button to press can be measured by the RT and the MT). Results indicated that RT and MT were positively correlated, and that the MT:RT ratios of the two ethnic groups were similar for elementary cognitive tasks but somewhat different for the Math Verification Test, which consists of single-digit addition, subtraction, and multiplication problems; the African-American children showed shorter MTs. Jensen argues that it is very unlikely that such strategy differences, if they exist and would be replicable, completely explain the correlations between g and performance differences. Because the evidence is derived entirely from studies involving elementary cognitive tasks, the generalization to more complex tests is not known.

There has been some debate in the literature as to whether SH reflects statistical artifacts. Some authors have argued that selecting two groups from a homogeneous population on the basis of their total test scores (as implicitly done in the comparison of African-Americans and European-Americans) inevitably leads to a confirmation of SH (e.g., Roskam & Ellis, 1992; Schönemann, 1992). However, it has been pointed out recently by Dolan (1997) that such a confirmation is not a mathematical necessity and can indeed only be expected under unrealistic sampling schemes. In a similar vein, Braden (1989) found a nonsignificant correlation between g loadings and the performance differences of deaf and hearing children.

Finally, Jensen has addressed test bias as an explanation of SH. He quotes evidence from a study by McGurk (1975), who found that EA—AA differences are larger for nonverbal than for verbal tests. This study refutes the argument that the style of language in tests, supposedly favoring European-Americans, contributes to performance differences. The most extensive study on the role of cultural factors in EA—AA differences has been reported by the same author (McGurk, 1951, 1953a, b; data were reanalyzed by Jensen & McGurk, 1987). A panel of "78 judges, including professors of psychology and sociology, educators, professional workers in counseling and guidance" (p. 295) were asked to classify items from well-known group-administered intelligence tests as "least cultural," "neutral," or "most cultural." The analyses revealed that the removal of presumably biased items did not affect the size of the observed B—W difference and that the item bias did not favor either statistical group. Also, the B—W differences were larger for the cultural than for the noncultural items

Unfortunately, McGurk's study suffers from two problems. The first has to do with the items that were used in the study. An inspection of the items that were rated as noncultural, such as verbal analogies, verbal opposites, and clock problems, suggests that at least some of the items indeed contain fairly strong cultural elements. Jensen and McGurk scrutinized the raters' implicit rationale for rating cultural loading; in their view cultural loading was mainly related to the "distinction between the recall of past-learned information and the mental manipulation of simple and familiar information that is provided in the test item itself" (p. 301). Clearly, the distinction between the recall of past information and reasoning is a poor rendering of cultural loading. A test

of the influence of cultural loading on ethnic performance differences requires a test battery without a confounding link between cultural loadings and mental transformations. The second problem is statistical. The authors tested item bias (differential item functioning), using an ANOVA of an item x race x subjects design. There are three difficulties with this design and analysis: (a) the dependent variable is dichotomous, which affects Type I and Type II error probabilities; (b) the analysis should be carried out per item (instead of for all items at once). In the design used by the authors the number of biased items is probably underestimated; (c) the authors should have added ability (sum score) as an additional independent variable (cf. Holland & Wainer, 1993; Lord, 1980; Mellenbergh, 1982; Van de Vijver & Leung, 1997). In sum, McGurk's study does not constitute an adequate test of the role of cultural loading as an alternative explanation of SH.

Studies of the Generalizability of SH

A few studies addressed the generalizability of SH to other ethnic groups. Lynn and Owen (1993), testing SH in South Africa among a group of Whites, Blacks, and Indians, found ambiguous results. The difference between the Whites and Blacks was not less than two standard deviations (SD) for 8 of the 10 subtests administered. The mean difference of Indians and Whites was one SD. The relationship between the Black g (i.e., the g loading as found in the factor analysis of the data of the Blacks) and performance difference differences between these groups was .62 ($p < .05$), thereby supporting SH. However, when the White g was used, no significant correlations were obtained. Similarly, the correlations between both the White and the Indian g and White—Indian differences were not significant.

Nagoshi, Johnson, DeFries, Wilson, and Vandenberg (1984) administered 15 mental tests to 5,333 Americans of Japanese, Chinese, and European ancestry. Of the six reported correlations between g loading and ethnic group differences, only two were significant.

Jensen and Whang (1994) studied performance differences among Chinese-Americans and African-Americans using Raven's Standard Progressive Matrices (as g measure) and 12 chronometric variables derived from the Math Verification Test. The Raven performances were significantly different for the groups (0.32 SD). The performances of the groups on the chronometric variables differed significantly (effect sizes for addition, subtraction, and multiplication were 0.47, 0.45, and 0.23, respectively) and these differences were related to g, but other factors seemed to be involved, too. The group differences in the chronometric tasks were larger than would be expected from the group difference in g. The Chinese pupils presumably had an advantage in speed of information processing, specifically the speed of retrieval of numerical information from memory possibly caused by extensive practice effects.

Finally, Te Nijenhuis and Van der Flier (1997) administered the GATB (General Aptitude Test Battery) tests to Dutch majority-group members ($n = 806$) and migrants ($n = 1332$), who on average lived 11.2 years in The Netherlands. The sample consisted of adults, mainly males, applying for blue-collar jobs at the Dutch Railways and region-

al bus companies. In comparison to majority-group members, these migrant groups have a lower level of mastery of the Dutch language, lower education levels, and are more often unemployed. Te Nijenhuis and Van der Flier found significant, positive correlations between g loadings (taken from the Dutch norms study) and standardized group differences in each group.

Towards a New Interpretation

Because there is so much evidence to support SH and there are so few studies that have successfully addressed alternative interpretations, SH seems to be unequivocally supported: "Since Spearman's hypothesis has been consistently borne out in many independent sets of appropriate data, and no contrary data have been found, it may legitimately claim the status of empirical fact" (Jensen, 1992, p. 232, and 1993, p. 48).

We question the validity of this conclusion and contend that a g loading often is not a pure measure of task complexity. A g loading may tap additional factors, such as knowledge of the language and culture of the test designer. Depending on the composition of the test battery, this "contamination" can be expected to be more salient in culturally more entrenched tests, particularly when the groups to be tested have a different knowledge of the linguistic and cultural background of the tests. In empirical studies employing common intelligence tests, the first principal component often confounds complexity and cultural and linguistic factors. Clearly, an adequate test of SH should disentangle these two components.

Differential mastery of the testing language by cultural groups creates a spurious correlation between g and intergroup performance differences if complex tasks require more linguistic skills than do simple tasks. A number of studies of SH have reported large g loadings for verbal tests (e.g., Peoples et al., 1995; Sandoval, 1982; Thorndike, Hagen, & Sattler, 1986). Similarly, in Carroll's (1993) model of cognitive abilities, crystallized intelligence, predominated by linguistic components, has a high loading on general intelligence. Tentative evidence for the influence of linguistic actors in testing SH can also be found in the earlier mentioned study by Te Nijenhuis and Van der Flier (1997). Their two samples, a Turkish-Dutch group (of first- and second-generation migrants) and a group of native Dutch, had undoubtedly mastered the testing language (Dutch) to different degrees. The score differences between the samples were regressed on g. The Vocabulary, Arithmetic Reasoning, and Computation tests had equally high g loadings (of about .7; see their Figure 1), but they revealed unequal group differences in performance levels. Vocabulary was .77 SD above the regression line while Arithmetic Reasoning and Computation were close to the regression line.

In the present study we examine the cultural loading of test material. Cultural loading is the generic term for implicit and explicit references to a specific cultural context, usually the culture of the test author, in the instrument or its administration (Van de Vijver & Poortinga, 1992). These loadings can create intergroup test score differences that are unrelated to the construct intended to be measured by the test. Like other

forms of bias, cultural loading is not an inherent property of an instrument but a characteristic of an intergroup comparison (Van de Vijver & Leung, 1997); the same test may yield valid differences in a comparison of Dutch and Belgian individuals, and may be affected by cultural loadings when comparing Dutch and British individuals. Cultural loadings can emanate from the stimulus medium, response medium and format, item content, and administration.

The present study examines SH in mental tests administered to majority-group and migrant primary school children in the Netherlands. An attempt is made to decompose g in verbal—cultural aspects and cognitive complexity. The relative contribution of complexity and verbal—cultural factors to observed cross-cultural performance differences are compared.

Method

Participants

A sample of 1221 primary school children, age 6 to 12 years, were selected from different regions in the Netherlands (the six- and seven-year old children were combined in the analyses). The sample consisted of Dutch majority-group members (n = 747), and a group of second-generation migrants (n = 474). In both cultural groups half were boys and half were girls. The majority of the participants were tested in urban regions where migrants mainly reside. In Table 1 the country of birth of the parents of the migrants is listed. Whereas the Surinamese and Antilleans make up 36% of the migrant population (Martens & Veenman, 1999), they only formed 11% of our sample; for Turks these figures are 26% and 22%, and for the Moroccans 22% and 36%, respectively. Not all migrant children speak Dutch when they enter school. The first language of Moroccan pupils is usually one of three Berber dialects or Arabic, while Turkish pupils speak Turkish (or in a few cases Kurdish) as their first language. Compared to Turks and Moroccans, children from Surinam and the Netherlands Antilles tend to have fewer language problems because of the relatively widespread usage of Dutch as home language.

Dutch is the language of education, except for some lessons in the native language and culture (about 2.5 hours per week). Special Islamic schools allot more time to learning the own language and culture. From these schools 75 pupils were tested in the study.

Table 1. *Country of Birth of the Migrants' Parents (Percentages; N = 474)*

	Percentage	
Birth country	Mother	Father
The Netherlands	7.9	4.4
Morocco	42.4	42.4
Turkey	29.3	29.7
Surinam or Netherlands Antilles	9.9	10.6
Elsewhere	10.4	12.5
Unknown		0.4

No first-generation migrants were involved in the study. First-generation migrants tend to have a lower level of knowledge of the Dutch language and culture than second-generation migrants. The inclusion of first-generation migrants would have boosted the influence of verbal and cultural aspects on intergroup performance differences. Restricting the study to second-generation children ensured that all children studied had followed a known (and across cultural groups equal) number of years of Dutch education, and had sufficient command of Dutch for the test administration. Yet, there is evidence that substantial differences in knowledge of the Dutch lexicon between majority-group pupils and migrant pupils linger on throughout the primary school period, even for second-generation children (Verhoeven, 2000).

Instruments

A computer-assisted cognitive ability test named the Tilburgse Allochtonen en Autochtonen Reactie Tijd Test (TAART) was administered. The test has been developed to assess simple cognitive processes, with little influence of cultural and linguistic knowledge (Helms-Lorenz & Van de Vijver, 1995; Van de Vijver & Willemse, 1991). It runs on IBM-compatible computers and uses the mouse as response device. The whole battery consists of nine subtests; results of the only two subtests that were administered to all age groups are reported here.

In Figure 1 geometric figures, as used in the items, are presented. In the first task (ECT1) five figures are shown, consisting of two pairs of identical stimuli and an "odd one out." The participant has to identify the latter. The second task (ECT2) involves "complementary figures." The figures c and d of Figure 1 are said to be complementary because they form exactly one black square when they are "added" (combined). Each ECT2 item consisted of two pairs of complementary figures and an "odd one out." The latter had to be identified by the pupil.

Both ECT1 and ECT2 consist of two series of ten items each, with a short break in between. When an item is presented on the screen, the mouse is located in the center of the screen in the "mouse box." This mouse box is surrounded by five squares, all at equal distance from the mouse box in a circular arrangement. The reaction time (used as performance measure) is defined as the time elapsed between stimulus onset and the moment the pupil moves the mouse outside the borders of the mouse box. In order to ensure that the pupil identifies the target figure before starting to move the mouse, the contents of the squares become gray and only the borders remain visible once the mouse leaves the mouse box. Pupils were instructed to respond fast without making any errors.

Both tests have four practice items. The computer gives feedback about correctness of responses (a face appears on the screen that is either happy or sad). The practice items are administered again if one or more incorrect responses are given. The actual testing starts when all four-exercise items have been solved correctly.

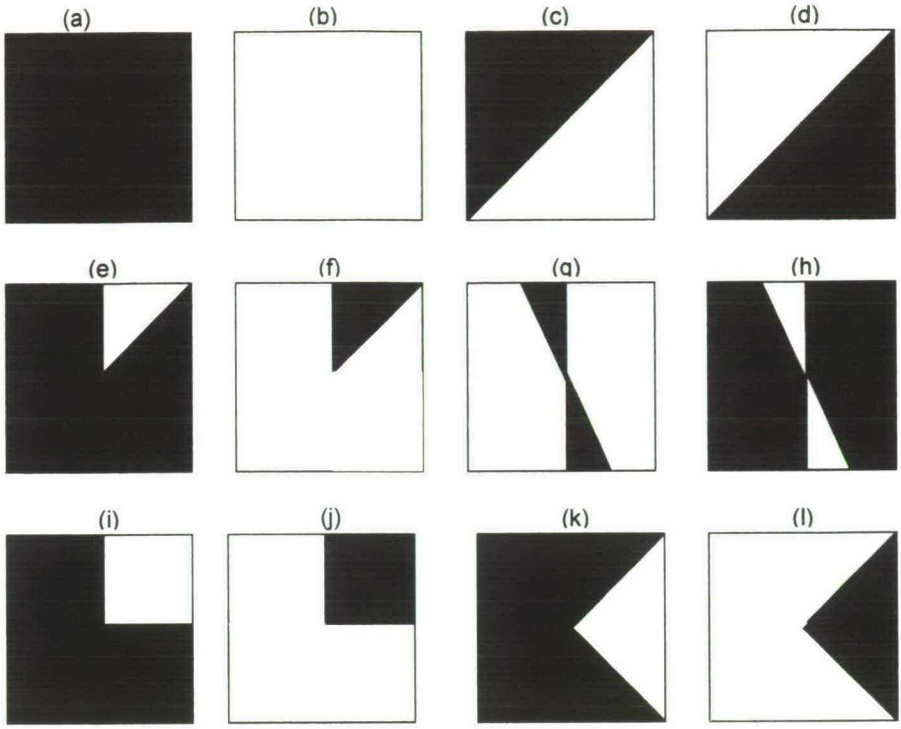


Figure 1. Geometric figures as usend in ECT's

Incorrect responses are treated as missing values in the data. In an analysis that is not further documented here, the proportion of errors was found to be small and similar for majority-group members and migrants. The internal consistencies of ECT1 and ECT2 (based on RTs) were .89 and .90, respectively.

Two individually administered intelligence tests were also administered: the Revised Amsterdamse Kinder Intelligentie Test (RAKIT) (Bleichrodt, Drenth, Zaal, & Resing, 1987) and the Revised Snijders-Oomen Nonverbal Intelligence Test (SON-R) (Laros & Tellegen, 1991). The reliability and validity of both tests have been shown in nation-wide samples; the test manuals provide age-specific norm tables. The COTAN, the committee that evaluates psychological tests in The Netherlands, gave favorable ratings to both tests (Evers, Van Vliet-Mulder, & Ter Laak, 1992). Furthermore, studies among migrant children have demonstrated the suitability of the tests for assessing these groups.

The SON-R (Laros & Tellegen, 1991) was originally intended for use with children that have a hearing impairment. Because the administration is nonverbal, it may also be an adequate test in populations with low proficiency in the testing language. Because of time constraints, a selection of four (out of seven) subtests were administered; Categories, Analogies (both abstract reasoning tests), Situations (concrete reasoning), and Mosaics (spatial relations).

Categories consists of three series of nine items, all in multiple-choice format. Three drawings of objects with a common characteristic (e.g., three different drawings of dogs) are given on a page. On the next page there are five drawings. The pupil has to point to the two drawings that belong to the same category (e.g., a mouse, a dog, a pencil, and building blocks).

Analogies uses geometric figures that are presented in the format a : b :: c : ?. The last figure has to be chosen from four alternatives depicted at the bottom of the page. The pupil has to discover the principle behind the change within the first pair of figures and apply it to the second pair; for example, the first figure is an empty square and the second figure is a square with a small black circle in its center. The third figure is an empty triangle. The four alternatives are: (i) triangle with a large empty circle in its center; (ii) triangle with a small black circle in the center; (iii) empty triangle; and (iv) a triangle with an empty small circle in the center. The test consists of three series of 11 items.

Situations, a multiple-choice test to assess concrete reasoning, also has three series of 11 items. Each item consists of a drawing with one or more missing parts. The correct solution has to be chosen from 4, 6, 9, or 10 alternatives. For example, the situation drawing can be a man walking with a leash in his hand; the object at the end of the leash is absent. The alternatives to choose from are a chicken, a dog, a frog, and a cat (each with a leash tied around its neck).

The Mosaics test is similar to Koh's Blocks used in the Wechsler scales. It is a performance test in which patterns are to be copied using white/red squares. The test has 20 items. Each mosaic pattern consists of nine fields and a field corresponds to one square. The pattern to be copied shows the colors but not the boundaries of the squares. The size of the stimulus pattern does not correspond to the size of the response pattern.

Not all children get the same items. After the example items have been presented, one always begins with the first item of the *a series* (10 items) and ends the series when the pupil has made two errors (not necessarily at successive items) or when the end of the series is reached. The administration of the *b series* (10 items) starts with the item number that is one less than the score of the *a series*. In order to estimate the internal consistency correct scores were assigned to items at the beginning of the test that had not been answered by the child, while zeros were assigned to all items not reached by the pupil. This treatment of missing value may lead to some overestimation of the tests' reliability. The internal consistencies of the current sample were .88 for Categories, .89 for Mosaics, .90 for Situations, and .92 for Analogies.

The short version of the RAKIT (Bleichrodt, Resing, Drenth, & Zaal, 1987) was administered, consisting of six tests. Exclusion is a multiple-choice test. The pupil has to choose one figure, among four abstract figures (the page is divided into four quadrants), that does not follow the rule applied to the other three figures. The test administration ends when the last item (50) is reached or when four successive items are solved incorrectly. The test measures logical reasoning, especially inductive thinking.

In Word Meaning, measuring active and passive vocabulary, a word is read aloud by the experimenter and from an array of four figures the pupil has to pick the one that depicts the word. For example, the verb "to read" is read out aloud by the administrator. Four figures presented in quadrants are a girl reading, a little girl phoning, an old lady knitting, and a toddler sleeping. The test administration ends when the last of the 60 items is reached or when four successive items are solved incorrectly.

Discs is a performance test that utilizes discs and a board with protruding pins. Three discs fit on each pattern of pins. The discs have two, three, or four holes and are to be placed over the corresponding pin formations. The pin patterns are arranged in three rows of three patterns each to accommodate nine discs. Depending on the age group, 12 or 18 discs are to be placed by the pupil. The discs are presented in two piles of nine discs each in a standardized sequence. The first disc for each set of pins is used for instruction purposes; and the remaining two positions are to be used by the child. This test is meant to measure pattern recognition and matching, speed and accuracy, eye-hand coordination, and spatial orientation.

Learning Names measures the ability to learn paired associates. The test booklet has 12 drawings of cats and butterflies. The pupil is shown a drawing while a name is read out aloud by the test administrator. Additional standardized cues are given in the form of an additional name or adjective to facilitate the learning process. The administrator reads the 12 names and shows each time the corresponding page of the booklet; the pupil is requested to remember the names. Then the pupil is asked to reproduce the name with each drawing. Feedback is given about the correctness of each response. The series is repeated. The number of items administered ranges from 2 x 10 to 2 x 12 depending on the age of the pupil.

Hidden Figures consists of a complex drawing depicted on the top half of a page. The bottom half of the page depicts six drawings. One of these six drawings forms part of the big drawing. The pupil is requested to identify the hidden pattern. The total number of items is 50; each age group starts at a different item. The test administration ends after 5 failures. This task requires visual analysis, pattern recognition, matching, and the ability to ignore distracting, irrelevant stimuli.

Finally, Idea Production has five test items. The pupil is asked to generate in a short, specified period of time as many words or names of objects or situations as possible, that belong to a broad category such as "things you can eat." The easier items at the beginning of the test are not given to the older age groups.

Our sample showed the following reliability coefficients: .82 for Exclusion, .89 for Idea Production, .80 for Learning Names, .67 for Discs, .79 for Hidden Figures, and .91 for Word Meaning.

Procedure

The administration time of ECT lasted 5 to 10 minutes per subtest. About half of the pupils completed the RAKIT and the other half the shortened version of SON-R. The SON-R took about 45 minutes and the RAKIT about 50 minutes to be administered.

Results

Test Characteristics

Measures of g Loadings. In line with Jensen's operationalization, g loadings of the tests were determined using principal component analysis. Because pupils completed the ECTs and either the RAKIT or SON-R, two separate analyses were needed. Standardized data, controlling for age, were factor analyzed per cultural group; one factor was extracted. The factor analysis of SON-R and ECTs produced eigenvalues of 2.67 for the majority-group members and 3.00 for the migrants, explaining 45% and 50% of the variance, respectively. The analysis for the RAKIT and ECTs revealed eigenvalues of 2.58 for the majority-group members and 2.50 for the migrants (i.e., explaining 32% and 31%). The agreement of the factor loadings in the majority-group and migrant sample was very high: Tucker's phi was .99 for the SON-R and ECTs, and .98 for the RAKIT and ECTs. These values strongly suggest factorial similarity in both cultural groups. In the remainder this factor is referred to respectively as, *migrants' and majority-group's* g; the mean of the two loadings is labeled *Jensen's* g. Factor loadings are presented in Table 2.

Table 2. *Complexity level, Carroll's and Jensen's g Loadings, Cultural Loading and Verbal Loading of Each Test*

	Measure							
			Maj. g	Maj. g	Mig. g	Mig. g		
Test	Complexity level[a]	Carroll's g[b]	SON-R	RAKIT	SON-R	RAKIT	Cultural loading[d]	Verbal loading[e]
				(a) RAKIT				
Word Meaning	4	7		.50		.54	4.03	130
Learning Names	3	6		.50		.55	2.83	242
Discs	4	5		.63		.65	1.24	97
Ideas	4	3		.39		.30	3.43	100
Hidden Figures	4	5		.72		.67	2.90	153
Exclusion	7	8		.57		.74	1.21	80
				(b) SON-R				
Analogies	8	8	.67		.81		1.34	78
Categories	7	8	.73		.75		3.83	56
Mosaics	4	5	.76		.77		1.72	41
Situations	4	5	.77		.75		3.97	60
				(c) ECT				
ECT1	3	1	.42	.53	.53	.41	1.72	75
ECT2	4	5	.44	.64	.57	.49	2.28	75

Note. Jensen's g is the mean of the g loading as found in the majority-group and in the migrant group. Maj. g = g factor as found in the data of the majority-group. Mig. g = g factor as found in the data of the migrant group

aDerived from Fischer's (1980) skill theory. bDerived from Carroll's (1993) "Structure of Cognitive Abilities". CDerived from factor analyses (loadings on the first factor). dBased on test ratings by 25 judges. eNumber of words in the test (instructions, test items, feedback, and response, as specified in the test manual.

Theoretically-Based Complexity Measures. A complexity measure of each test was based on Carroll's (1993) model of the structure of cognitive abilities (p. 626), which synthesizes existing factor-analytic work. The order of the lower-order factors in the model ranks the strength of their relationship with g (p. 625). The first factor, fluid intelligence, has the strongest and the last (eighth) factor, processing speed, has the weakest relationship with g. Rank order numbers were used as theoretically based complexity ratings (see Table 2) and are referred to as *Carroll's* g.

A second measure was based on a theoretical analysis of complexity rules. Intra-test complexity rules, usually based on cognitive process analysis, have been discussed by various authors (e.g., Laros & Tellegen, 1991; Pellegrino & Glaser, 1979; Schorr, Kiernan, & Bower, 1981; Spelberg, 1987; Tanzer, Gittler, & Ellis, 1995). However, to our knowledge, no theoretical analyses have been conducted to determine intertest complexity rules. Therefore, we relied on Fischer's (1980) Skill Theory, which is a neo-Piagetian model of cognitive development. According to the theory, children develop skills of gradually increasing complexity. Skills can be broken down into elementary building blocks. Ten developmental levels of increasing skill complexity are postulated. Skills of a lower level are combined to form new, more complex skills, thus forming hierarchical levels. These levels are divided into three tiers: sensory-motor actions, representations, and abstract skills. In the Appendix a brief summary is given of the levels as well as the rationale for the complexity level assigned to each of the tests used in the present study. The score assigned to a test corresponds to the minimal developmental level needed to accomplish the task, and is used as a rank order measure of task complexity (see Table 2). The scoring was done jointly by the authors (the scoring was deemed to be too complex for raters unfamiliar with Skill Theory).

Verbal Loading. Verbal loading was operationalized as the total number of words in the instructions, test material presented to the pupil, pupil's response (i.e., the number of core terms for scoring as specified in the test manual), and feedback, including words used for explaining the task or encouraging the pupil (see Table 2).

Measure of Cultural Loading. The cultural loading of all tests was rated by 25 third-year psychology students, who had followed at least two courses in cross-cultural psychology. The ratings were gathered in two sessions. In the first session cultural loadings of the tests were rated on a scale of 0 to 5 (0 = none, 1 = very low, 2 = low, 3 = moderate, 4 = high, and 5 = very high). Cultural loading was defined for the raters as "the extent to which the test contains cultural elements." A score of zero had to be assigned if no cultural elements were judged to be present in the test (i.e., the test could be applied to all cultural groups without adaptations). Each test was rated individually. During the second session, a week later, the items were rated. Figure tests were not rated at item level because the items of these tests do not use stimuli that vary in cultural loading.

The means of the cultural loading ratings of each test are given in Table 2. The overall interrater reliability (internal consistency) was .94; the intraclass correlation was

.88. The reliability of the test level ratings was .86 (intraclass correlation: .72) and of the means derived from the item level ratings .89 (intraclass correlation: .85). Correlations between ratings for tests and items were larger than .90 for all tests. In conclusion, the interrater agreement was good.

Item- and the test-level ratings were combined in a single analysis; item-level ratings were averaged per test. The cultural loadings were factor analyzed, using an Oblimin rotation (delta = 0). A solution with three factors could well be interpreted (eigenvalues: 10.09, 3.32, and 1.94, together explaining 73% of the variance). The first factor represents knowledge of the Dutch culture, involving the verbal and non-verbal tests that were rated as requiring much cultural knowledge (e.g., Idea Production, Categories, and Situations) (see Table 3). The second factor is mainly defined by the two computer tests; the factor was labeled computer mode. The figure tests showed the highest loadings on the third factor, which was called figure mode. The correlations of the factors were positive (first and second: .19; first and third .49: second and third: .16).

Table 3. *Factor Loadings of the Three Factors Derived from an Oblimin Factor Analysis on the Cultural Loading Ratings*

	Factor		
Stimulus	Culture	Computer mode	Figure mode
		Item-level ratings	
RAKIT			
Word Meaning	.79	-.01	.17
Learning Names	.57	.07	.34
Idea Production	.78	.13	-.16
Hidden Figures	.50	-.04	.39
SON-R			
Analogies	.03	.19	.74
Categories	.66	-.01	.29
Situations	.89	-.09	.10
		Test-level ratings	
RAKIT			
Word Meaning	.69	-.31	.28
Learning Names	.13	-.16	.54
Discs	.27	-.02	.70
Idea Production	.96	.19	-.27
Hidden Figures	.33	-.02	.64
Exclusion	-.11	.12	.90
SON-R			
Analogies	-.05	.15	.91
Categories	.74	.08	.09
Mosaics	.21	.24	.42
Situations	.88	-.05	.06
ECT			
ECT1	.18	.89	-.23
ECT2	.02	.92	.19

Aggregate Measures. Jensen's g loading, the two complexity ratings, and verbal loadings were factor analyzed, together with the three raters' factors. Two factors were extracted, with eigenvalues of 3.31 and 1.87, explaining 74% of the variance. An Oblimin rotation (delta = .10) was carried out. Carroll's g, Jensen's g, figure mode, and complexity (derived from the Skill theory) constituted the first factor (see Table 4). The high loadings of the figure tests is not surprising, because the figure tests employed, Analogies and Exclusion, have a high cognitive complexity. The factor is labeled "aggregate g." Cultural loading and verbal loading showed a high positive loading on the second factor while computer mode showed a strong, negative loading. The factor is labeled "aggregate c" (c for culture).

Table 4. *Rotated Factor Loadings of the Second Order Factor Analysis (Pattern Matrix)*

	Factor	
Measure	Aggregate g	Aggregate c
Complexity[a]	.87 (.88)	-.26 (-.31)
Carroll's g[b]	.86 (.86)	.22 (.18)
Jensen's g[c]	.83 (.81)	-.12 (-.05)
Figure mode[d]	.80 (.78)	.33 (.31)
Cultural factor[d]	.06 (.11)	.74 (.73)
Computer mode[d]	-.39 (-.41)	-.85 (-.84)
Verbal loading[e]	-.33 (-.34)	.72 (.73)

Note. Values between parentheses refer to loadings after correction for attenuation of Jensen's g.
aDerived from Fischer's (1980) skill theory. bDerived from Carroll's (1993) Structure of Cognitive Abilities. cDerived from factor loadings on first common factor (majority-group and migrants combined). dDerived from factor in ratings by students. eNumber of words in the test.

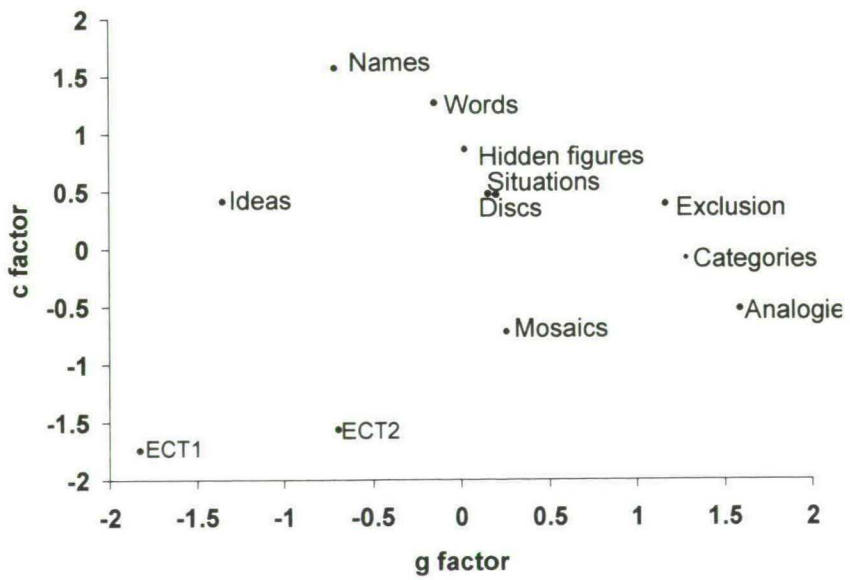


Figure 2. Factor scores of the tests

The correlation between aggregate g and aggregate c was low (.08 before and .06 after correction for attenuation). This low correlation and the absence of high secondary loadings of the measures demonstrate that g and c were well distinguishable in the present battery. That the relationship was low can also be seen in the scatter plot of the factor scores in Figure 2.

Table 5. *Sample Means, Standard Deviations and Sample Sizes for the Migrants and for the Majority-group per Age Group*

		Cultural group					
		Migrants			Majority-group		
	Age	M	SD	N	M	SD	N
Test		(a) RAKIT					
Word	6-7	31.94	7.11	18	38.06	3.50	32
Meaning	8	39.52	6.38	21	43.26	3.46	31
	9	40.40	3.91	15	45.58	4.32	33
	10	44.00	4.72	21	48.10	5.37	29
	11	47.64	6.14	22	53.33	4.52	33
	12	45.36	3.86	14	52.37	4.46	30
Learning	6-7	10.11	1.97	18	13.64	4.09	33
Names	8	12.57	3.63	21	16.00	3.79	31
	9	13.33	2.77	15	17.39	3.19	33
	10	15.05	4.30	21	18.93	3.47	29
	11	16.64	4.02	22	19.76	3.66	33
	12	14.93	4.23	14	18.67	2.88	30
Discs	6-7	289.83	74.57	18	218.53	64.26	32
	8	245.48	72.25	21	188.65	72.62	31
	9	191.20	66.33	15	168.97	57.08	33
	10	193.62	67.31	21	137.00	40.35	29
	11	161.00	64.11	22	139.45	49.20	33
	12	166.36	46.38	14	143.20	48.82	30
Idea	6-7	42.78	11.74	18	52.00	15.48	32
Production	8	54.43	20.20	21	62.42	16.83	31
	9	58.07	20.53	15	65.21	15.08	33
	10	61.24	14.93	21	76.86	17.28	29
	11	72.27	19.45	22	80.67	15.98	33
	12	72.43	23.45	14	83.03	23.70	30
Hidden	6-7	25.39	4.73	18	28.61	4.57	31
Figures	8	32.76	4.04	21	31.81	4.61	31
	9	33.00	4.46	15	32.64	5.24	33
	10	34.24	3.32	21	36.66	3.19	29
	11	37.95	3.21	22	38.33	3.26	33
	12	37.43	2.82	14	38.57	2.96	30
Exclusion	6-7	29.28	5.77	18	31.75	5.74	32
	8	37.67	4.62	21	35.61	5.75	31
	9	36.20	9.02	15	38.79	5.50	33
	10	36.67	5.08	21	40.90	4.24	29
	11	42.95	4.15	22	43.33	4.13	33
	12	42.64	2.34	14	43.30	5.52	30

(Table continues)

Table 5. (continued)

		Cultural group					
		Migrants			Majority-group		
	Age	M	SD	N	M	SD	N
Test				(b) SON-R			
Analogies	6-7	10.59	4.72	32	11.84	4.61	44
	8	13.86	4.91	22	14.97	5.78	34
	9	16.00	4.71	29	17.53	4.95	36
	10	15.81	3.19	16	19.51	4.78	45
	11	20.04	5.03	23	21.34	5.44	35
	12	18.86	5.78	22	20.32	5.30	38
Categories	6-7	8.03	3.43	32	8.50	3.41	44
	8	10.41	3.54	22	11.59	4.53	34
	9	11.45	3.82	29	13.42	3.95	36
	10	12.53	5.05	17	14.78	4.20	45
	11	15.09	4.04	23	15.86	3.27	35
	12	14.82	4.50	22	15.34	4.23	38
Mosaics	6-7	7.72	3.17	32	9.61	3.56	44
	8	10.14	3.98	22	11.41	3.39	34
	9	12.03	3.38	29	13.33	4.41	36
	10	11.76	3.09	17	14.60	3.60	45
	11	14.17	3.07	23	15.26	3.25	35
	12	14.32	3.87	22	15.76	3.29	38
Situations	6-7	12.09	4.70	32	14.25	4.60	44
	8	13.95	4.67	22	16.09	4.96	34
	9	17.24	5.08	29	18.17	4.49	36
	10	16.24	4.42	17	20.13	4.86	45
	11	20.96	4.58	23	20.86	4.88	35
	12	19.23	4.82	22	20.53	4.30	38
				(c) ECT			
ECT1	6-7	1879.32	668.17	80	1709.27	466.99	130
	8	1607.56	465.57	67	1478.90	490.45	112
	9	1545.74	410.75	87	1440.00	328.99	122
	10	1513.46	535.40	83	1375.25	326.67	113
	11	1374.44	364.25	81	1226.65	266.62	113
	12	1399.44	418.32	69	1159.83	269.69	87
ECT2	6-7	6413.52	1531.36	69	6130.28	1642.86	126
	8	5627.94	1508.05	66	5326.14	1353.99	107
	9	4851.28	1110.99	85	4524.20	1125.52	122
	10	4513.55	875.50	83	4095.68	1010.97	113
	11	3663.68	751.50	81	3304.38	658.26	112
	12	3784.17	957.36	69	3284.64	754.91	87

Note: Higher scores refer to better performances, except for Discs and ECTs in which shorter reaction times (lower scores) refer to better performance

It could be argued that a factor analysis is not allowed on these data, as some data are rank orders. However, a multidimensional scaling procedure yielded dimensions quite similar to the factors described.

Performance Differences

In Table 5 the sample means, standard deviations, and sample sizes are listed per age group for migrants and majority-group members.

Two MANOVAs of the test data were used to test the effects of culture (two levels), gender (two levels), and age (six levels); separate analyses of the intelligence tests were necessary because no participants had taken all subtests (Table 6). Ten out of 12 subtests showed a significant main effect for culture ($p < .05$); majority-group members invariably obtained higher scores. The RAKIT showed the largest ethnic differences; culture explained on average 11% of the variance; for the SON-R ands ECTs these figures were 4% and 1%. Main effects for age were found for all tests ($p < .01$), with older pupils showing better performance. Age effects were larger than culture and gender effects, explaining on average 33% of the variance. Two tests (Word Meaning and Mosaics) revealed a main effect for gender ($p < .05$); both showed higher scores for males. Overall, however, gender differences were small, explaining on average less than 1%. A few univariate interactions were significant; these are not further considered because the effects were neither large nor of primary interest here.

Table 6. *Multivariate Analysis of Variance Testing the Effects Culture, Gender and Age and Their Proportion of Variance Explained (h^2).*

	Main effects					
	Culture		Gender		Age	
Test	F^a	η^2	F^a	η^2	F^a	η^2
			RAKIT			
Word Meaning	86.07**	.24	6.25*	.02	60.09**	.52
Learning Names	75.74**	.22	1.67	.01	18.76**	.25
Discs	33.97**	.11	2.91	.01	19.58**	.26
Idea Production	22.34**	.08	.40	.00	20.11**	.27
Hidden Figures	3.30	.01	.79	.00	45.61**	.45
Exclusion	4.77*	.02	.26	.00	29.20**	.35
			SONR			
Analogies	12.40**	.03	3.56	.01	27.16**	.28
Categories	8.57**	.02	2.58	.01	26.56**	.27
Mosaics	17.08**	.05	11.63**	.03	30.18**	.30
Situations	14.51**	.04	.32	.00	20.62**	.23
			ECT			
ECT1[b]	3.14/2.09	.01/.01	1.66/.06	.01/.00	26.11**/12.70**	.27/.19
ECT2[b]	6.86**/1.65	.02/.00	.01/.11	.00/.00	66.08**/40/34**	.49/.42

adf = 1, 348. bFirst number in cell of ECT1 refers to ECT–RAKIT group, the second to the ECT–SON-R group. *p < .05. **p < .01.

Table 7. *Effect Sizes per Age Group, averaged over age groups and the averaged effect size corrected for attenuation*

	Age (in years)							
Test	6 and 7	8	9	10	11	12	Mean[a]	Corrected mean[b]
				(a) RAKIT				
Word Meaning	-1.20	-.77	-1.23	-.80	-1.09	-1.64	-1.12	-1.23
Learning Names	-1.01	-.92	-1.32	-1.01	-.82	-1.11	-1.03	-1.29
Discs	-1.05	-.78	-.37	-1.06	-.39	-.48	-.69	-1.03
Idea Production	-.65	-.44	-.42	-.96	-.48	-.45	-.57	-.64
Hidden Figures	-.70	.22	.07	-.75	-.12	-.39	-.28	-.35
Exclusion	-.43	.39	-.38	-.92	-.09	-.14	-.26	-.32
				(b) SON-R				
Analogies	-.27	-.20	-.32	-.84	-.25	-.27	-.36	-.39
Categories	-.14	-.28	-.51	-.51	-.21	-.12	-.29	-.33
Mosaics	-.56	-.35	-.33	-.82	-.34	-.41	-.47	-.52
Situations	-.47	-.44	-.20	-.82	.02	-.29	-.36	-.41
				(c) ECT				
ECT1	-.31	-.27	-.29	-.32	-.48	-.70	-.39	-.44
ECT2	-.18	-.21	-.29	-.44	-.51	-.59	-.37	-.42

Note: Negative effect size points to higher performance of majority-group pupils.
aMean effect size. bMean effect size corrected for attenuation (divided by test reliability; for the SON-R manual values were used and for ECT and RAKIT sample values were used.

Correlations between Test Characteristics and Effect Sizes

Effect sizes, defined as the difference of majority-group members and migrants divided by their pooled standard deviation, are presented in Table 7.

Correlations are reported between effect sizes and various test characteristics: empirical g measures (majority-groups', migrants', and Jensen's g), theoretical complexity measures (Carroll's g and Fischer's complexity), the three raters' factors (cultural factor, computer mode, and figure mode), and verbal loading. Correlations were computed for two types of effect sizes; first, the effect sizes averaged over age groups were used in the correlations (referred to in Table 8 as "averaged data," $n = 12$); furthermore, each age group was treated as an independent replication, thereby constituting 72 observations (6 age groups x 12 tests) ("unaveraged data"). As can be seen in Table 8, the averaged and unaveraged data yielded a largely similar pattern of findings; the major difference was the smaller number of significant correlations for the averaged data, due to the small sample size. For the averaged data, only verbal loading ($r = .67$) and the aggregate c factor ($r = .65$) showed significant correlations ($p < .05$). Culturally more entrenched tests showed larger performance differences. For the unaveraged data, all empirical g measures and complexity ratings showed negative correlations with effect sizes ($p < .01$), with the exception of a nonsignificant correlation of Carroll's g. The aggregate g factor showed a significant, negative correlation of -.24 ($p < .05$) with effect size. The sign of these correlations is negative, indicating that, contrary to Jensen's studies (e.g., 1993), *smaller* performance differences were found for tests with

higher g loadings. Correlations of effect sizes with the raters' factors were weaker; the only significant correlation was found for the computer mode in the unaveraged data (r = -.29, p < .05). Finally, verbal loadings showed significant correlations, both averaged (r = .67, p < .05) and unaveraged (r = .67, p < .01); higher verbal loadings give rise to more performance differences between majority-group members and migrants.

The correlations suggest that ethnic performance differences were stronger related to culture than to cognitive complexity. The issue was further explored in a multilevel regression analysis, with items as level-1 units and tests as level-2 units (Bryk & Raudenbush, 1992; Goldstein, 1987). The two factors were the independent variables explaining the effect size. The slopes were held fixed while the intercept was allowed to vary randomly across classes and tests. The regression coefficient was .11 for the g factor (ns) and .18 (p < .001) for the c factor. So, the multilevel analysis conformed that the c factor was more important than the g factor in explaining ethnic performance differences in this data set.

In sum, the prediction from SH that the intergroup differences in cognitive performance would increase with the tasks' g loading was not borne out; on the contrary, performance differences decreased with increasing g loadings. Verbal and cultural loading had a salient impact on effect size; differences in cognitive test performances between migrants and majority-group members increased with these loadings. Clearly, the data do not support SH.

Table 8. *Correlations between Effect Sizes of 12 Tests and g, Cultural Loading, Task Complexity, and Verbal Loading, Both for the Six Age Groups Separately ("Unaveraged", Based on 6 Age Groups x 12 Tests, and "Averaged", Combining All Age Groups)*

Measure	Correlation	
Empirical g measures	Unaveraged (N = 72)	Averaged (N = 12)
Migrants' g[a]	-.30**	-.37
Majority-group's g[b]	-.36***	-.45
Jensen's g[c]	-.34**	-.41
Cognitive complexity measures		
Carroll's g[d]	.02	.03
Complexity[e]	-.35**	-.48
Raters' factors		
Cultural factor[f]	.21	.26
Computer mode[f]	-.29*	-.41
Figure mode[f]	.01	-.10
Verbal loading[g]	.67*	.67*
Aggregate measures		
Aggregate g[h]	-.24*	-.28
Aggregate c[h]	.65*	.65*

aLoadings on first factor in migrants' data. bLoadings on first factor in majority-groups' data. cDerived from factor loadings on first common factor (majority-group and migrants combined). dDerived from Carroll's (1993) Structure of Cognitive Abilities. eDerived from Fischer's (1980) skill theory. fFactors in ratings by students. gNumber of words in the test (instructions, test items, feedback, and response, as specified in the test manual). hAggregate g and c factors (see Table 4) *p < .05.

Discussion

SH was tested in a sample of Dutch majority-group and second-generation migrant pupils (aged 6 to 12 years), using two widely applied intelligence tests and a computer-assisted reaction time test. The common operationalization of g as the loading on the first factor was deemed inadequate to test SH because it confounds cognitive complexity and verbal—cultural loading. An attempt was made to disentangle these two components. Theoretically based measures of cognitive complexity were derived from Carroll's (1993) model of cognitive abilities and Fischer's (1980) skill theory. Cultural loadings of tests were assessed by ratings of the test materials by 25 senior psychology students. The verbal loading of a test was operationalized as the number of words in the test. A factor analysis of all test aspects revealed two oblique factors, g and c. There was tentative evidence that c was at least as important as cognitive complexity in the explanation of performance differences of majority-group and migrant children.

Our results are at variance with common findings in the literature on SH. The major departure involves the failure to find a positive contribution of cognitive complexity to the prediction of cross-cultural performance differences. Two possible explanations can be envisaged to explain the discrepancy. The first involves the composition of the test battery. It could be argued that the tests employed in the present study are poorly suited for testing SH. In our view this argument is implausible. The test battery was composed of both elementary cognitive transformations and more common cognitive tests in order to obtain a broad coverage of the intellectual domain. Furthermore, the tests used in this study were selected to minimize bias effects. All tests chosen in the present study were originally designed for multicultural groups and attempt to assess cognitive skills with a minimal reliance on acquaintance with the Dutch language and culture. Finally, an adequate test of SH assumes that g and c are unrelated, as was the case in our data. Looking at common intelligence tests, one cannot escape from the impression that the g—c relationship will often be positive because tests that require extensive verbal processing (these may even involve figure tests) are often the cognitively more complex tasks in intelligence tests. This introduces a spurious, positive relation between cognitive complexity and verbal processing, which complicates the interpretation of g loadings and challenges their adequacy to test SH.

Second, it could be argued that the external validity of the present findings is limited to the Netherlands or possibly to Western Europe and that results cannot be generalized to comparisons of African-Americans and European-Americans. Although some characteristics of the migrant groups studied are specific to Western Europe, such as the high prevalence of Mediterranean, Islamic groups, other characteristics are common to various minority groups, such as a lower level of education, SES, income, and higher level of unemployment than the majority group (Martens & Veenman, 1999). The samples studied here have the underprivileged position shared by many recently migrated groups. Moreover, the IQ difference of about 1 SD that is often found between African-Americans and European-Americans is not far from the difference of 0.7 SD for

the SON-R and 1.1 SD for the RAKIT of the present study. As an aside it may be noted that the larger differences on the RAKIT may be related to the more salient verbal and cultural aspects of the RAKIT as compared to the SON-R.

In sum, our instruments and samples offered an adequate framework for testing SH that is not too dissimilar from the North American context in which most tests of SH took place. It remains to be determined in future studies to what extent the prominent role of cultural factors in the explanation of performance differences is replicable. The present study clearly underscores the need to "purify" g measures and to disentangle cognitive complexity and cultural entrenchment in tests of SH.

References

Bleichrodt, N., Drenth, P. J. D., Zaal, J. N., & Resing, W. C. M. (1987). *RAKIT Handleiding, Revisie Amsterdamse Kinder Intelligentie Test.* Lisse, the Netherlands: Swets & Zeitlinger.

Braden, J. P. (1989). Fact or artifact? An empirical test of Spearman's hypothesis. *Intelligence, 13,* 149-155.

Bryk, A. S., & Raudenbush, S. W. (1992). *Hierarchical linear models: Applications and data analysis.* Newbury Park, CA: Sage.

Carroll, J. B. (1993). *Human cognitive abilities. A survey of factor-analytic studies.* Cambridge: Cambridge University Press.

Dolan, C. (1997). A note on Schönemann's refutation of Spearman's Hypothesis. *Multivariate Behavioral Research, 32,* 319-325.

Evers, A., Van Vliet-Mulder, J. C., & Ter Laak, J. (1992). *Documentatie van Tests en Testresearch in Nederland.* Nederlands Instituut voor Psychologen (NIP), Amsterdam.

Fischer, K. W. (1980). A theory of cognitive development: The control and construction of hierarchies of skills. *Psychological Review, 87,* 477-531.

Goldstein, H. (1987). *Multilevel models in educational and social research.* London: Griffin.

Helms-Lorenz, M., & Van de Vijver F. J. R. (1995). Cognitive assessment in education in a multi-cultural society. *European Journal of Psychological Assessment, 11,* 158-169.

Helms-Lorenz, M., & Van de Vijver F. J. R. (in preparation). *TAART Test Manual.*

Holland, P. W., & Wainer, H. (Eds.) (1993). *Differential item functioning.* Hillsdale, NJ: Erlbaum.

Humphreys, L. G. (1985). Race differences and the Spearman hypothesis. *Intelligence, 9,* 275-283.

Jensen, A. R. (1982). Reaction time and psychometric g. In H. J. Eysenck (Ed.), *A model for intelligence.* Springer.

Jensen, A. R. (1984). The Black–White difference on the K-ABC: Implications for future tests. *Journal of Special Education, 18,* 377-408.

Jensen, A. R. (1985). The nature of Black–White difference on various psychometric tests: Spearman's hypothesis. *Behavioral and Brain Sciences, 8,* 193-263.

Jensen, A. R. (1992). The importance of intraindividual variation in reaction time. *Personality and Individual Differences, 3,* 925-928.

Jensen, A. R. (1993). Spearman's Hypothesis tested with chronometric information processing tasks. *Intelligence, 17,* 47-77.

Jensen, A. R. (1998). *The g factor. The science of mental ability.* Westport, CT: Praeger.

Jensen, A. R., & McGurk, F. C. J. (1987). Black—white bias in "cultural" and "non-cultural" test items. *Personality and Individual Differences, 8,* 295-301.

Jensen, A. R., & Reynolds, C. R. (1982). Race, social class, and ability patterns on the WISC-R. *Personality and Individual Differences, 3,* 423-438.

Jensen, A. R., & Whang, P. A. (1994). Speed of accessing arithmetic facts in long-term memory: A comparison of Chinese-American and Anglo-American children. *Contemporary Educational Psychology, 19, 1-12.*

Laros, J. A., & Tellegen, P. J. (1991). *Construction and validation of the SON-R 5 ½- 17, the Snijders-Oomen non-verbal intelligence test.* Groningen, the Netherlands: Wolters-Noordhoff.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems.* Hillsdale, NJ: Erlbaum.

Lynn, R., & Owen, K. (1993). Spearman's hypothesis and test score differences between Whites, Indians, and Blacks in South Africa. *Journal of General Psychology, 121,* 27-36

Martens, E. P., & Veenman, J. (1999). De sociaal-economische positie van etnische minderheden. In H. M. A. G. Smeets, E. P. Martens, & J. Veenman (Eds.), *Jaarboek minderheden* (pp. 107-138). Houten, the Netherlands: Bohn Stafleu Van Loghum.

McGurk, F. C. J. (1951). *Comparison of the performance of negro and white high school seniors on cultural and non-cultural psychological test questions.* Washington, DC: The Catholic University of America Press.

McGurk, F. C. J. (1953a). On white and negro test performance and socioeconomic factors. *Journal of Abnormal and Social Psychology, 48,* 448-450.

McGurk, F. C. J. (1953b). Socioeconomic status and culturally-weighted test scores of Negro subjects. *Journal of Applied Psychology, 37,* 276-277.

McGurk, F. C. J. (1975). Race differences - twenty years later. *Homo, 26,* 219-239.

Mellenbergh, G. J. (1982). Contingency table models for assessing item bias. *Journal of Educational Statistics, 7,* 105-118.

Montie, J. E., & Fagan, J. F. (1988). Racial differences in IQ: Item analysis of the Stanford-Binet at 3 years. *Intelligence, 12,* 315-332.

Naglieri, J. A., & Jensen, A. R. (1987). Comparison of Black–White differences on the WISC-R and the K-ABC: Spearman's hypothesis. *Intelligence, 11,* 21-43.

Nagoshi, C. T., Johnson, R. C., DeFries J. C., Wilson, J. R., & Vandenberg, S. G. (1984). Group differences and principal-component loadings in the Hawaii Family Study of Cognition: A test of the generality of "Spearman's hypothesis." *Personality and Individual Differences, 5,* 751-753.

Pellegrino, J. W., & Glaser, R. (1979). Cognitive correlates and components in the analysis of individual differences. *Intelligence,* 3, 187-214.

Peoples, C. E., Fagen, J. F., & Drotar, D. (1995). The influence of race on 3-year-old children's performances on the Standford-Binet fourth edition. *Intelligence, 21,* 69-82.

Roskam, E. E., & Ellis, J. (1992). Commentary on Guttman: The irrelevance of factor analysis for the study of group differences. *Multivariate Behavioral Research, 27,* 205-218.

Sandoval, J.(1982). The WISC-R factorial validity for minority groups and Spearman's hypothesis. *Journal of School Psychology, 20,* 198-204.

Schorr, D., Kiernan, R., & Bower, G. (1981). *Analysis versus synthesis in block design tests.* Stanford: Stanford University.

Schönemann, P. H. (1992). Extensions of Guttman's results from g to PC1. *Multivariate Behavioral Research, 27,* 219-223.

Spearman, C. (1927). *The abilities of man.* New York: Macmillan.

Spelberg, H. C. L. (1987). *Grenzen testen.* Groningen, the Netherlands: Foundation of Child Studies.

Stankov, L. (1983). The role of competition in human abilities revealed through auditory tests. *Multivariate Behavioral Research.* Monograph no. 83-1.

Tanzer, N. K., Gittler, G, & Ellis, B. B. (1995). Cross-cultural validation of item complexity in a LLTM-calibrated spatial ability test. *European Journal of Psychological Assessment, 11,* 170-183.

Te Nijenhuis, J., & Van der Flier, H. (1997). Comparability of GATB scores for immigrants and majority group members: Some Dutch findings. *Journal of Applied Psychology, 82,* 675-687.

Thorndike, R. L., Hagen, E. P., & Sattler, J. M. (1986). *Stanford-Binet Intelligence Scale* (4th Ed.). Riverside, CA: DLM Teaching Resources.

Van de Vijver, F. J. R. (1997). Meta-analysis of cross-cultural comparisons of cognitive test performance. *Journal of Cross-Cultural Psychology, 28,* 678-709.

Van de Vijver, F. J. R. (1999). Inductive Reasoning in Zambia, Turkey, and The Netherlands: Establishing Cross-Cultural Equivalence. (under review)

Van de Vijver, F. J. R., & Leung, K. (1997). Methods and data analysis of comparative research. In J. W. Berry, Y. H. Poortinga, & J. Pandey (Eds.), *Handbook of cross-cultural psychology* (2nd ed., vol. 1, pp. 257-300). Boston: Allyn & Bacon.

Van de Vijver, F. J. R. & Poortinga, Y. H. (1992). Testing in culturally heterogeneous populations: When are cultural loadings undesirable? *European Journal of Psychological Assessment, 8,* 17-24.

Van de Vijver, F. J. R., & Willemse, G. R. (1991). Are reaction time tasks better suited for ethnic minorities than paper-and-pencil tests? In N. Bleichrodt & P. J. D. Drenth (Eds.), *Contemporary issues in cross-cultural psychology* (pp. 450-464). Lisse, the Netherlands: Swets & Zeitlinger.

Verhoeven, L. (2000). Detectie van taalachterstand. In N. Bleichrodt & F. J. R. Van de Vijver (Eds.), Diagnostiek bij allochtonen (pp. 180-204). Lisse, the Netherlands: Swets & Zeitlinger.

Vernon, P. A., & Jensen, A. R. (1984). Individual and group differences in intelligence and speed of information processes. *Personality and Individual Differences, 5,* 411-423.

Appendix
Rationale for determination of cognitive complexity

This section provides a brief summary is given of Skill Theory levels as well as the rationale for the complexity level awarded to each of the tests used in the present study. The reader is referred to Fischer (1980) for a comprehensive description of the theory.

Sensory-Motor tier: Levels 1 to 4
Skills are composed of Sensory-Motor (S-M) sets.
Level 1
Level one is characterized by undifferentiated, uncoordinated, multimodal sets. Examples of a sensory-motor sets: 'to look at a doll' and 'to grasp a doll'. No differentiation between for example sight and sound, both modalities can be mixed in one set.
Level 2: Sensory-Motor Mapping
A S-M set is mapped onto another S-M set. Example: to use 'looking at the doll' to guide person to 'grasp the doll'. So one action is used to bring about another action.
Level 3
S-M system. The components of one set are related to the components of another set. Example. Child watches bread falling then watches bread crumb falling and comes to understand relationship between variations in what is dropped to variations in falling behavior. The limitations of this developmental level are that the child can only use one system at a time, and cannot understand objects independently of their own actions.

Repesentational tier
Level 4
Two S-M systems are combined to form a representational set. Child is capable of representing simple properties of objects or events and people independently of their own actions. This level is characterized by lack of differentiation. Example the child confuses effects of weight and size especially if these covary.
Level 5: Representational Mapping
One representational set is mapped onto another representational set. Example: large weight causes large stretch of a spring, small weight causes small spring stretch. Therefore weight determines length of stretch.
Level 6: Representational System
Subsets of one representational set is related to subsets of another.
Limitations: Child can deal with one system only, cannot relate systems to one another. Cannot think of objects in the abstract.

Abstract Tier
Level 7: Abstract Set
Child can abstract intangible attributes that characterize broad categories. It can con-

trol two representational systems simultaneously. It understands how changes covary without manipulating objects.

Level 8

Abstract mapping. Child is able to relate 2 abstract sets.

Level 9: Abstract System

Child is able to produce flexible and differentiated relation between abstract concepts.

Level 10:

Person is able to coordinate different abstract systems.

Task analysis:

Fischer (1980) provides guidelines for task analysis. He demonstrates the application of the guidelines on the development of social roles, not for cognitive tasks. The task analysis of our cognitive measures are based on these guidelines. Fischer notes that "Even with these guidelines, doing a task analysis is no trivial matter. Unfortunately, it still involves a degree more art than I would like" (p. 506).

Five guideline questions:

1. Does the skill require sensory-motor, representational, or abstract sets?
2. What are the sources of variation that the person must control in the skill?
3. What are the relations between sets that the child must control?
4. What is the particular task, and what must the person control to perform it?
5. What is the minimal task that would demonstrate the skill in question? (irrelevant complexities?).
6. In the Appendix the answers per guideline question per test can be found.

ECT:

Task 1: 4 identical, 1 odd man out

1. Does the skill require sensory-motor, representational, or abstract sets?

*Sensory-Motor: **Level 3.** Two sensory motor sets are combined to form a S-M system. Components of set 1: blocks and to grasp; set 2: identical blocks are clustered. The clustering entails visual matching, in skill theory terms: the sensory component is the visual stimulus and the action is to cluster identical stimuli. The subject is required to understand that it can guide itself by looking at a block to grasp it, and that identical figures go together. The subject must be able to cluster identical figures mentally and to pinpoint the odd-man out.*

4. What is the particular task, and what must the person control to perform it?

Five squares are presented. Four are identical and one is different from the rest. The subject is required to match (group identical figures by means of visual matching) objects mentally. Four blocks are identical and one is different. The subject must identify the object that does not belong to the rest. This is an exclusion task. So the subject must control grasping, grouping together and identifying the odd-one out.

5. What is the minimal task that would demonstrate the skill in question? (irrelevant

complexities?) *The minimal task would be to supply the subject with a group of squares that are very different from the odd-one out. The exclusion format might introduce irrelevant complexity. The stimulus & response medium (computer) might introduce irrelevant complexity. This task has no irrelevant task complexities.*

Task 2: 2 pairs complementary, 1 odd man out

1. Does the skill require sensory-motor, representational, or abstract sets?

Representational, ***Level*** *4. Two S-M systems are combined to form a representational set. The sorting system is combined with the complementary system. The sorting system has been described for task 3. The complementary system consists of the principle of finding a 'good fit' (as a key fits only into a specific key whole). The sensory object is grasped and fitted into the correct place. The subject understands how changes covary without manipulating objects.*

4. What is the particular task, and what must the person control to perform it?

The task is to mentally combine two complementary squares, twice, to remember their positions, and to pick the square left over. The subject must control the concept of complements, must control remembering positions and must locate the odd one out and must be able to indicate this my means of the mouse.

5. What is the minimal task that would demonstrate the skill in question? (irrelevant complexities?) *The minimal task would have been a set of three (not 5) figures to demonstrate the complement concept (level 3). This task entails the additional concept of categorization. This combination seems to be the minimum task to illustrate this level 4 skill.*

SON-R

1. Categories

1. Does the skill require sensory-motor, representational, or abstract sets?

Abstract, ***Level*** *7. Subject understands how changes covary without manipulating objects. The subject is able to abstract intangible attributes that characterize broad categories*

4. What is the particular task, and what must the person control to perform it?

The subject is presented 3 figures and is required to deduce the underlying similarity of this set that forms a category. The subject is requested to pick 2 figures, from an array of 5 pictures, which belong to the implied category. The remaining 3 figures do not belong to the implicit category.

5. What is the minimal task that would demonstrate the skill in question? (irrelevant complexities?) *The minimal task would entail a clear category, and the three figures that do not belong to the category should be quite obvious. This might not always be the case, adding irrelevant complexity. Categories of semantic content are not clear-cut cross-culturally.*

2. Mosaics

1. Does the skill require sensory-motor, representational, or abstract sets?

Representational. ***Level*** *4. A level 3 S-M system building a one-dimensional figure is combined with a second S-M system of building in a second dimension forming a representational set where the two become related. It is not necessary to understand these figures that are to be built in an*

abstract way, because they can manipulate the tiles manually (not mentally). A child will be able to reconstruct the figure without understanding covariations of the two dimensions in an abstract way.

4. What is the particular task, and what must the person control to perform it?

The subject is presented with a picture and is requested to use tiles to reconstruct the picture. The format is 3x3, 9 tiles.

5. What is the minimal task that would demonstrate the skill in question? (irrelevant complexities?) *The minimum task would have been a 2x2 two-dimensional format. (A 2x2 format would however not allow as much increase in complexity as a 3x3 format allows.)*

3. Situations

1. Does the skill require sensory-motor, representational, or abstract sets?

Representational. ***Level 4****. The child can represent simple properties of objects, events and people independently of own actions (for example; a man walking with a leash in his hand, implies a dog at the end of the leash). They do not need to relate this set to another nor understand this in an abstract way.*

4. What is the particular task, and what must the person control to perform it?

The subject is presented with a picture which illustrates a situation, for example; a lady looking in a mirror. One (or more) part(s) of the picture is (are) missing. The object is to find the missing parts amongst a number of alternatives.

5. What is the minimal task that would demonstrate the skill in question? (irrelevant complexities?) *A simple pair of objects that logically go together should measure this skill. The more complex situations however, do not only measure transformations and developments of this skill, they introduce irrelevant complexities.*

4. Analogies

1. Does the skill require sensory-motor, representational, or abstract sets?

Abstract. ***Level 8*** *The subject is able to abstract an intangible rule that is applied in an example and understands the changes depicted without manipulating the objects. This rule is then mapped onto another set of stimuli. Child is able to relate 2 abstract sets.*

4. What is the particular task, and what must the person control to perform it?

In the top part of the item an example is given. The subject is requested to deduce the applied rule and to apply the rule to the item below. Four alternatives are presented of which the subject is to choose the correct one.

5. What is the minimal task that would demonstrate the skill in question? (irrelevant complexities?) *I do not think that irrelevant complexities are present.*

RAKIT

Exclusion

1. Does the skill require sensory-motor, representational, or abstract sets?

Abstract, ***Level*** *7. Subject understands how changes covary without manipulating objects. The subject is able to abstract intangible attributes that characterize broad categories.*

4. What is the particular task, and what must the person control to perform it?
The subject is presented 4 geometrical figures that consist of squares, triangles, circles, lines, etc. The subject is requested to search for three figures that go together and to pinpoint the figure that does not go with the rest.
5. What is the minimal task that would demonstrate the skill in question? (irrelevant complexities?) *The exclusion format might introduce irrelevant complexity. It might have been more straightforward to request the subject to indicate which figures go together. Three figures might also have been sufficient to illustrate the skill (less complexity increase possible?).*

Word meaning

1. Does the skill require sensory-motor, representational, or abstract sets?
Representational. ***Level 4****. The subject is able to associate an object with a given word.*
4. What is the particular task, and what must the person control to perform it?
A word is read to the subject and four pictures are presented to the subject depicted on a single page in a booklet. The subject is requested to pinpoint the picture associated to the word.
5. What is the minimal task that would demonstrate the skill in question? (irrelevant complexities?). *Maybe one picture should rather have been presented and the subject should have been requested to produce the word. This test may contain irrelevant complexity for subjects of a minority culture.*

Discs

1. Does the skill require sensory-motor, representational, or abstract sets?
Representational, ***Level 4****. The child can represent the concept of a pattern of wholes fitting over pins arranged in the same pattern (as is a complementary task), independently of their own actions.*
3. What are the relations between sets that the child must control?
The child must relate an object can fit into another object and that this combination is unique.
4. What is the particular task, and what must the person control to perform it?
The subject is presented a board with protruding pins. The pins are arranged in 3 patterns arranged in 3 rows. Then the subject is given a number of discs with wholes in them (fixed order). The object is to place the disc over the appropriate protruding pins as fast as possible. This is a speeded test. The subject is timed per disc. The score entails the time, measures in seconds, required to place the disc correctly.
5. What is the minimal task that would demonstrate the skill in question? (irrelevant complexities?) *I don't think there is any irrelevant complexity. The speededness of the test might be irrelevant*

Learning names

1. Does the skill require sensory-motor, representational, or abstract sets?
Representational, ***Level 3****. The subject is able to associate an object with a nonsensical name, as all people have names. The names are existing words but have no semantic connection with the*

object to be named. This test is at a lower level than test "word meaning" because it has less or no semantic connotation, and it requires more short term memory than long-term memory.

4. What is the particular task, and what must the person control to perform it?

The child is presented with a picture (ex. a cat) and is read a nonsensical name simultaneously. The child is requested to remember the name the animal or insect is called on the picture. A series of 12 pictures is presented with names, and the procedure is repeated once. Then the child is to name the series of pictures independently, twice.

5. What is the minimal task that would demonstrate the skill in question? (irrelevant complexities?). *The names to be learned are less nonsensical, than intended. This introduces irrelevant complexity to the children unfamiliar to the (unintended) semantic connotation.*

Hidden figures

1. Does the skill require sensory-motor, representational, or abstract sets?

This kind of test is believed (Fischer, 1980) to be a test for object permanence. The object permanence skill is a sensory-motor set (skill). This task measures a more developed sense of object permanence, the representation of it. This ***level 4*** *skill requires the child to relate object permanence independently of their own actions.*

4. What is the particular task, and what must the person control to perform it?

The subject is presented with a drawing that contains various complete and incomplete objects and lines that overlap. No colors are used. Below the drawing 6 pictures of objects are presented. The subject is requested to pinpoint the objects that corresponds 100% (size, angle and completeness) with a part of the big drawing.

5. What is the minimal task that would demonstrate the skill in question? (irrelevant complexities?) *The stimulus material is considered to be too complex and therefore it is not seen as a minimal task to measure object permanence. The distracters are too similar to the response alternatives. Lets say the items are too tricky, and misleading. The instructions are too short and allow for unmeant errors.*

Idea Production

1. Does the skill require sensory-motor, representational, or abstract sets?

Representational. ***Level 4****. The child can represent simple properties of objects, events and people independently of own actions (for example; the child understands that only small objects can fit into a pocket). They do not need to relate this set to another nor understand this in an abstract way.*

4. What is the particular task, and what must the person control to perform it?

The subject is to imagine all possible objects that for example might be found in a pocket of a jacket. The subject must name as many objects he/she can think of in a short period of time (1 minute). The number of words generated is used as score. 5 Situations are sketched: 1. What can one drink? 2. What can one pick up? 3. Where can one hide? 4. What can one find in a shop? What can one do in the street?

5. What is the minimal task that would demonstrate the skill in question? (irrelevant

complexities *"What can one do in the street" is most likely to be an inappropriate item for a Moslem girl, who is not permitted to play in the street. "What can one drink" might elicit less responses from Moslem children whose parents do not drink any form of alcohol. These factors add irrelevant complexities to the task*

Chapter 3

An Empirical Study of Bias in Culture-Reduced Tests: Its Detection and Antecedents

Michelle Helms-Lorenz
Fons J.R. van de Vijver

Abstract

Construct, method, and item bias were studied in a sample of majority-group members (n = 679), and first (n = 232) and second (n = 471) generation migrants in The Netherlands. The subjects were 6 to 12 years old. Twelve culture-reduced subtests derived from two standardized intelligence batteries, as well as two tasks from a computer-assisted elementary cognitive test battery were used. Exploratory factor analytic solutions of subtest scores obtained in both cultural groups were compared to assess construct bias. The analysis of method bias, based on migrant data, estimated the influence of non-cognitive participant characteristics (acquaintance with the Dutch culture and with computers) on test performance. Item bias was assessed using both logistic regression and ANOVA. Both construct and item bias could be identified in some of the tests used. Method bias was found in almost all the tests used. Both participant- and test characteristics predicted unconditional bias, but did not predict conditional bias.

Introduction

The aim of this study is to explore the role of bias in performance on culture-reduced cognitive tests. Van de Vijver and Leung (1997) define bias in cross-cultural research as "all nuisance factors threatening the validity of cross-cultural comparisons" (p. 10). They distinguish between three types of bias. *Construct bias* occurs when the construct measured is not identical across cultural groups. Factor analysis is often used to compare the structure underlying an instrument in different cultural groups. *Method bias* is a generic name for all sources of cross-cultural score differences attributable to test characteristics (e.g., stimulus familiarity), samples (e.g., differential education or motivation), or test administration (e.g., lack of standardization, tester effects) (cf. Mercer, 1984). *Item bias* or *Differential Item Functioning* (DIF) refers to measurement artifacts at item level.

Two types of statistical techniques have been proposed to detect item bias. In conditional procedures score levels (ability groups) are taken into account in the detection of item bias. Item bias occurs "if individuals with equal ability but from different groups do not have the same probability of answering an item correctly" (Shepard, Camilli, & Averill, 1981, p. 319). Many techniques have been proposed to detect item bias, such as Logistic Regression (Swaminathan, Hambleton, & Rogers, 1989), Analysis of Variance (Van de Vijver & Leung, 1997), the Mantel-Haenszel Statistic (Holland & Wainer, 1993), and Item Response Theory (Hambleton & Van der Linden, 19978). Mellenbergh (1982) has proposed a distinction between two types of item bias. Uniform bias refers to nuisance factors influencing scores to the same degree at all score levels, while nonuniform bias refers to influences that are not identical across score levels. Conditional item bias techniques are popular in the area of cognitive and educational testing.

In unconditional procedures item bias is identified without any split in score levels. Exploratory factor analysis is the most frequently employed technique to study

unconditional item bias. This technique is commonly employed in the study of personality (e.g., Chan, Ho, Leung, Cha, & Yung, 1999; Eysenck, Barrett, & Eysenck, 1985; McCrae & Costa, 1997). Unconditional procedures are often applied in studies that focus on construct identity (structural equivalence; Van de Vijver & Leung, 1997) often performed in the field of personality research. Conditional procedures focus on the comparability of test scores often found in the field of mental testing. A brief (and necessarily selective) review of bias studies is presented.

(a) Construct Bias Studies

Studies among Western samples have found ample evidence for the structural equivalence of mental tests in schooled populations (for reviews see e.g., Irvine, 1979; Jensen, 1980; Van de Vijver, 1997).

(b) Method Bias Studies

The presence of method bias cannot be derived from the administration of an instrument in two cultural groups; its presence can only be demonstrated in a guided search assessing the impact of specific indicators on test performance, such as differential previous test exposure.

Sample characteristics. Foorman, Yoshida, Swank, and Garson (1989) administered the Raven (and other tests) to Japanese and American pupils of the same grades in their respective countries. No differences in general cognitive abilities were found between the two studied groups. They found that accuracy rates improved from second to fifth grade, and for the American pupils response latencies correspondingly increased. Japanese children's error rates decreased too, but this was accompanied by relatively little latency increase between the ages of 7 and 10. They showed that these samples had differential speed-accuracy tradeoffs due to differences in cultural styles. This expeditious Japanese response style was attributed to persistent training. Similar findings have been reported by Smith and Caplan (1988).

Test-wiseness can be an important source of performance differences on cognitive and educational tests (Rogers & Yang 1996). Despite the obvious relevance of this notion for cross-cultural research, there are almost no relevant studies. Van de Vijver, Daal, and Van Zonneveld (1986) carried out a training study of inductive thinking among upper primary school children in the Netherlands, Surinam, and Zambia. The latter group had no experience with mental testing. The Zambian group gained more from training than the Dutch and Surinamese groups, for example in a task in which vowels were to be identified. This was ascribed to differential test experience.

Test attitude is another sample characteristic believed to influence cognitive performance. In one of the studies by Arvey, Strickland, Drauden, and Martin (1990) a Test Attitude Survey (TAS), as well as three employment tests, were administered to 223 Anglo-Americans and 64 African Americans job applicants. They found that part of the ethnic performance differences were accounted for by TAS scores.

In a meta-analysis of cross-cultural performance differences, Van de Vijver (1997) reported a significant correlation of .37 between national differences in affluence and

performance. Interestingly, a similar correlation (of .39) was observed when the analysis was restricted to studies dealing with simple tasks. It appears that even cognitively simple tasks have characteristics that give rise to cross-national performance differences. It is tempting to conclude that, in line with the work on test-wiseness, these country differences are due to familiarity with stimuli, response procedures, and testing situations in general.

Instrument characteristics. Both stimulus and response features have been scrutinized. In particular stimulus familiarity can have a strong influence on cross-cultural differences. Deregowski and Serpell (1971) asked Scottish and Zambian children to sort miniature models of animals and motor vehicles and in another condition to sort photographs of these models. No cross-cultural differences were found for the actual models, while the Scottish children obtained higher scores than the Zambian children when photographs were sorted.

A study by Serpell (1979) demonstrates the effect of stimulus/response familiarity on test performance. Zambian and British children were asked to reproduce certain patterns using paper and pencil as well as flexible wire. The Zambian children scored higher in wire modeling (a popular pastime in Zambia) and lower in drawing than their English age mates, entirely in line with what could be expected from a response familiarity model.

Chan and Schmitt (1997) explored the degree to which response mode (test method) can reduce subgroup differences while keeping test content (and test constructs) constant. They administered a situational judgment test using a video-based and a paper-and-pencil method to 113 African- and 128 Anglo-American psychology undergraduates. In line with the authors' expectation, the latter method favored Anglo- over African-Americans to a greater extent than the former method.

(c) Item Bias Studies

Scheuneman (1979) analyzed the item pool of the Metropolitan Readiness Tests (1976 version) to identify biased items for Anglo-American and African-American groups in the USA. The test consists of language items, auditory items and visual items. Within the set of biased items, significantly more items were from the language area. A subset of the language items involved negative structures, such as "Mark the thing that is unopened" or "Mark the picture which shows neither a cat nor a dog". Of the biased language items, 86% involved negative forms.

It has been argued by J. Helms (1992) that cognitive ability tests fail to assess intelligence adequately, because they fall short in accommodating the emphasis on social relationships and the effect of social context on reasoning in the African-American culture. DeShon, Smith, Chan, and Schmitt (1998) investigated the effect of social context on reasoning in this cultural group. Wason conditional reasoning items were administered to test whether a social form of the items would diminish performance differences between Anglo- and African-Americans. Contrary to expectation their results did not confirm the hypothesis that items embedded in social context would show smaller performance differences.

Schmitt (1988) examined responses of Anglo-Americans? Whites (N = 278,166), Mexican-American (N = 2,963) and Puerto Rican (N = 3,230) candidates to the verbal part of Scholastic Aptitude Test Verbal test. She found that true cognates (words with a common root in English and Spanish) and items of special interest for Hispanic participants enhanced their performance.

Despite their impressive number and often high level of psychometric sophistication, item bias studies have not advanced our insight in the reasons underlying this bias. In Bond's (1993) words: "Theories about why items behave differently across groups can be described only as primitive" (p. 278); or in Linn's (1993) words: "The majority of items with large DIF values seem to defy explanation of the kind that can lead to more general principles of sound test development practice" (p. 359). The only item characteristic that shows a fairly consistent association with item bias is item difficulty (e.g., Linn, 1993): More difficult items tend to show more bias.

Item bias is the most extensively studied type of bias. This focus on items as the source of bias has two undesired, related consequences. First, anomalies at a global instrument level, such as differences in stimulus or response familiarity, are unlikely to be retrieved in an item bias analysis. Second, item bias has almost become synonymous to bias. As a consequence, there are almost no studies in which more than a single type of bias has been examined, although there are no theoretical reasons to assume that the most important sources of bias affect only single items or small subsets of items. By considering more than a single level of bias, a more comprehensive picture of all instrument-related problems may be obtained. Problems of item bias studies, such as the low cross-sample stability of bias findings (e.g., Holland and Wainer, 1993) may be easier to overcome by examining bias in a broader framework.

Present Study

The Netherlands has become a multicultural society in the last decades. Enrollment of first- and second-generation pupils in all types of education has necessitated the development of new or adaptation of existing educational and cognitive tests (Bleichrodt & Van de Vijver, 2000). The present study aims to detect construct, method, and item bias in different cognitive and educational measures that were designed for use in the Netherlands and to identify some of the antecedents.

Method

Participants

A sample of 1382 primary school children, age 6 to 12 years, were selected from different regions in the Netherlands (the six- and seven-year old children were combined in the analyses). Half were boys and half were girls. The sample consisted of Dutch majority group members ($\underline{n}$ = 679), and first ($\underline{n}$ = 232) and second ($\underline{n}$ = 471) generation migrants (see Table 1). The latter two groups were born, or had parents who were born in Morocco (38%), Turkey (25%), Surinam/the Netherlands Antilles (13%), or elsewhere outside the Netherlands (24%). The majority of the participants were

recruited from urban regions where migrants mainly reside. The sample of migrants is not fully representative of the national population. The Surinamese and Antillean groups are underrepresented as they make up 36% of the total migrant population (Martens & Veenman, 1999). The Turkish figure is fairly appropriate (26% of the migrant population), while Moroccan youngsters who make up 22% of the migrant population, are over represented.

Table 1. *Number of Participants per Culture (Migrants and Majority Group Members) and Age Group*

Age (yrs)	First-generation	Second-generation	Majority
6-7	25	83	130
8	45	67	114
9	36	87	122
10	49	84	113
11	35	81	113
12	42	69	87

Often migrant children hardly speak Dutch when they enter school. The first language of Moroccan pupils is usually one of three Berber dialects or Moroccan-Arabic, while Turkish pupils speak Turkish (or Kurdish in a few cases) as their first language. Compared to Turks and Moroccans, children from Surinam and the Netherlands Antilles tend to have fewer language problems because of the relatively widespread usage of Dutch as their home language.

Dutch is the language of instruction, except for some lessons in the native language and culture (about 2.5 hours per week). Special Islamic schools, visited mainly by Turkish and Moroccan pupils, allot more time to instruction in the own language and culture. From these schools 75 pupils were tested in the study. Differences in mastery of the Dutch language (the testing language) can be expected between the majority and migrant groups. There is evidence that substantial differences in knowledge of the Dutch lexicon, the most important source of linguistic differences between the ethnic groups continue throughout the primary school period (Verhoeven, 2000). These differences are more prominent in the first generation, but are still clearly present in the second generation.

Instruments

Two subtests of a computer-assisted cognitive ability test battery named the Tilburgse Allochtonen en Autochtonen Reactie Tijd Test (TAART) were administered. The test was developed to assess simple cognitive processes, with little item-specific influence of cultural and linguistic knowledge (Helms-Lorenz & Van de Vijver, 1995; Van de Vijver & Willemse, 1991). TAART is computer based, using the mouse as response device. The whole battery consists of nine subtests; results of the only two subtests that were administered to all age groups are reported here.

In the first task (ECT1) five figures are shown, consisting of two pairs of identical stimuli and an "odd one out" (see Figure 1), which the participant has to identify. The second task (ECT2) involves "complementary figures." Two figures are said to be complementary when they form exactly one black square when they are "added" (combined) (see Figure 1 for an example). Each ECT2 item consisted of two pairs of complementary figures and an "odd one out." The latter had to be identified by the pupil.

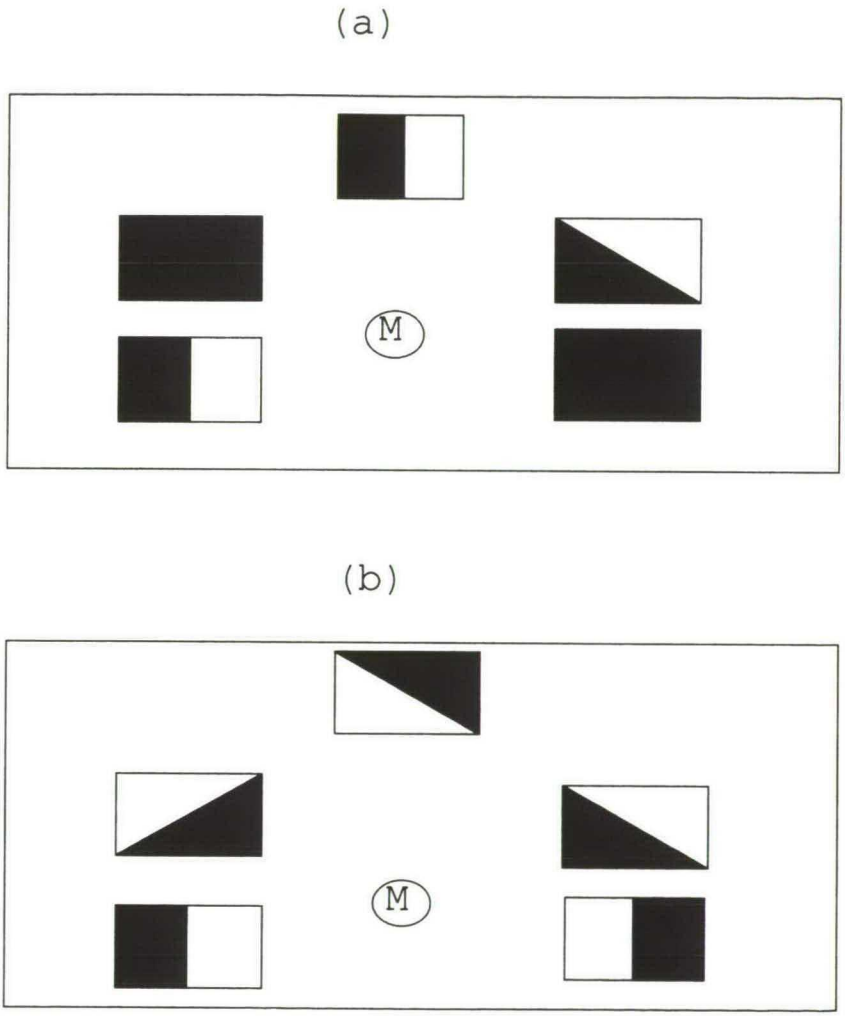


Figure 1 Example of ECT1 (panel a) and ECT2 (panel 2)

Note. In ECT1 there are two pairs of equal figures and an "odd-one-out" (the right upper figure). In ECT2 there are two pairs of complementary figures (i.e., they form exactly one square when put on top of each other) and an "odd-one-out" (the upper left figure). "M" indicates the box where the mouse is located at the beginning of an item.

Both ECT1 and ECT2 consist of two series of ten items each, with a short break in between the series. When an item is presented on the screen, the mouse is located in the center of the screen in a small square, the "mouse box." This square is surrounded by five equidistant squares, in which the figures appear. The reaction time (used as performance measure) is defined as the time (in ms) elapsed between stimulus onset and the moment the pupil moves the mouse outside the borders of the mouse box. In order to ensure that the pupil identifies the target figure before starting to move the mouse, the contents of the squares become gray and only the borders remain visible once the mouse leaves the mouse box. Pupils were instructed to respond fast without making any errors.

Both tests have four practice items. The computer gives feedback about correctness of responses (a face appears on the screen that is either happy or sad). The practice items are administered again if one (or more) incorrect response is given. The actual testing starts when all four exercises have been solved correctly.

Incorrect responses are treated as missing values in the data. An analysis of variance (per age group) was run on the numbers of incorrect responses. Some age groups showed a significant main effect for culture for both ECT1 and ECT2. The migrant pupils made more errors. On average 6.61% more majority group members completed the tasks without errors compared to the migrant group members. See Table 2 for the results. As can be seen from the Eta Squared values, these differences were small. Even though the migrants made more errors, the reaction times of these incorrect responses followed the same pattern for both cultural groups (see Helms & Van de Vijver, in preparation). These results reject the speed-accuracy trade-off interpretation for differences in error rates. The internal consistencies of ECT1 and ECT2 (based on RTs) were .89 and .90, respectively.

Table 2. *Analysis of Variance of Numbers of Incorrect Responses Per Age Group Testing the Effect for Culture, and the Proportion of Variance Explained by Culture (h^2).*

	ECT1		ECT2	
AGE GROUP	F	η^2	F	η^2
6 and 7 YEARS	5.72*	.02	3.57	.02
8 YEARS	.19	.00	1.28	.01
9 YEARS	2.91	.01	8.96**	.04
10 YEARS	11.36**	.04	1.85	.01
11 YEARS	6.21*	.03	7.22**	.03
12 YEARS	.00	.00	2.08	.01

*$p < .05$. **$p < .01$.

Two intelligence batteries were also administered (individually): the Revised Amsterdamse Kinder Intelligentie Test (RAKIT) (Bleichrodt, Drenth, Zaal, & Resing, 1984) and the Revised Snijders-Oomen Nonverbal Intelligence Test (SON-R) (Laros & Tellegen, 1991). The reliability and validity of both tests have been shown in nation-

wide samples; the test manuals provide age-specific norm tables. The COTAN, a standing committee of the Dutch Psychological Association that evaluates psychological tests in The Netherlands, gave favorable ratings to both tests (Evers, Van Vliet-Mulder, & Ter Laak, 1992).

The SON-R (Laros & Tellegen, 1991) was originally intended for use among children with a hearing impairment. Because the administration is nonverbal, it might well prove to be also an adequate test in populations with low proficiency in the testing language. Because of time constraints four of seven subtests were administered: Categories, Analogies (both abstract reasoning tests), Situations (concrete reasoning), and Mosaics (spatial relations).

Categories consist of three series of nine items and three practice items, all in multiple-choice format. Three drawings of objects with a common characteristic (e.g., three different drawings of dogs) are given on one page. On the next page there are five drawings. The pupil has to point to the two drawings that belong to the same category (e.g., a mouse, a dog, a pencil, a pile of building blocks, and another dog).

Analogies uses geometric figures that are presented so that a : b :: c : ?. The last figure has to be chosen from four alternatives depicted at the bottom of the page. The pupil has to discover the principle behind the change within the first pair of figures and apply it to the second pair. For example, the first figure is an empty square and the second figure is a square with a small black circle in its center. The third figure is an empty triangle. The four alternatives are: (i) triangle with a large empty circle in its center; (ii) triangle with a small black circle in the center; (iii) empty triangle; and (iv) a triangle with an empty small circle in the center. The test consists of three series of 11 items and 3 practice items.

Situations, a multiple-choice test tapping concrete reasoning, also has three series of 11 items. Each item consists of a drawing with one or more missing parts. The correct solution has to be chosen from 4, 6, 9, or 10 alternatives. For example, the situation drawing can be a man walking with a leash in his hand; the object at the end of the leash is absent. The alternatives to choose from are a chicken, a dog, a frog, and a cat (each with a leash around its neck).

The Mosaics test is similar to Koh's Blocks used in the Wechsler scales. It is a performance test in which patterns are to be copied using white/red squares within a specified amount of time. The test has two series of 10 items and 3 practice items. Each mosaic pattern consists of nine fields and a field corresponds to one square. The pattern to be copied shows the colors but not the boundaries of the squares. The size of the stimulus pattern does not correspond to the size of the response pattern.

Not all children get the same items. After the examples have been presented, the administration always begins with the first item of the *a series* and ends the series when the pupil has made two errors (not necessarily at successive items) or when the end of the series is reached. The administration of the *b series* starts with the item number that is one less than the score of the *a series*. In order to estimate the internal consistency cor-

rect scores were assigned to items at the beginning of the test that had not been answered by the child (items are ordered in difficulty order), while an incorrect score was assigned to all items not reached by the pupil. This treatment of missing value leads to some overestimation of the tests' reliability. The internal consistencies of the current sample were .88 for Categories, .89 for Mosaics, .90 for Situations, and .92 for Analogies.

The short version of the RAKIT (Bleichrodt et al.,1984) was administered, consisting of six tests. Exclusion is a multiple-choice test. The pupil has to choose one figure, among four abstract figures (the page is divided into four quadrants), that does not follow the rule applied to the other three figures. The test administration ends when the last item (50) is reached or four successive items are solved incorrectly. The test measures logical reasoning, especially inductive thinking.

In Word Meaning, measuring active and passive vocabulary, a word is read aloud by the experimenter and from an array of four figures the pupil has to pick the one that depicts the word. For example, the verb "to read" is read out aloud by the administrator. Four figures presented in quadrants are a girl reading, a little girl phoning, an old lady knitting, and a toddler sleeping. The test administration ends when the last of the 60 items is reached or when four successive items are solved incorrectly.

Discs is a performance test that utilizes discs and a board with protruding pins. Three discs fit on each pattern of pins. The discs have two, three, or four holes and are to be placed over the corresponding pin formations. The pin patterns are arranged in three rows of three patterns each to accommodate nine discs. Depending on the age group, 12 or 18 discs are to be placed by the pupil. The discs are presented in two piles of nine discs each in a standardized sequence. The first disc for each set of pins is used for instruction purposes, while the remaining two positions are to be used by the child. The test measures pattern recognition and matching, speed and accuracy, eye-hand coordination, and spatial orientation. The score is the number of seconds needed to place the discs in the right position.

Learning Names measures the ability to learn paired associates. The test booklet has 12 drawings of cats and butterflies. The pupil is shown a drawing while a name is read out aloud by the test administrator. Standardized cues are given in the form of an additional name or adjective to facilitate the learning process. The administrator reads the 12 names and shows each time the corresponding page of the booklet; the pupil is requested to remember the names. The pupil is then asked to reproduce the name of each drawing. Feedback is given about the correctness of each response. The series is repeated. The number of items administered ranges from 2 x 10 to 2 x 12 depending on the age of the pupil.

Hidden Figures consists of a complex drawing depicted on the top half of a page. The bottom half of the page depicts six drawings. One of these six drawings forms part of the big drawing. The pupil is requested to identify the hidden pattern. The total number of items is 50; each age group starts at a different item. The test administra-

tion ends after 5 failures. This task requires visual analysis, pattern recognition, matching, and the ability to ignore distracting, irrelevant stimuli.

Finally, Idea Production has five test items. The pupil is asked to generate in a short, specified period of time as many words or names of objects or situations as possible, that belong to a broad category such as 'Things you can eat'. Participants of all the age groups are presented with the same five items. The test is explained by means of one practice item. Correct responses are added together to form a total score per item. The sum total of the five items forms the test score.

Our sample showed the following internal consistency coefficients: .82 for Exclusion, .89 for Idea Production, .80 for Learning Names, .67 for Discs, .79 for Hidden Figures, and .91 for Word Meaning.

Procedure

The administration of the ECT tasks lasted five to ten minutes per subtest. About half of the pupils completed the RAKIT and the other half the SON-R. The SON-R took about 45 minutes and the RAKIT 45-60 minutes to be administered.

CITO tests. The oldest pupils in primary school usually participate in so-called CITO tests (nation wide administered school achievement measures, called after the CITO, the Centraal Instituut voor Toets Ontwikkeling, which develops the tests). The administration of the tests is not compulsory for schools, but the CITO test scores provide important input to the advice given to the pupil concerning secondary school choice. Test scores on the CITO Information, Language, and Arithmetic tests were available for most of the older participants of the present study (n = 130, of which 63 were of the majority group).

Grade marks. Three grade marks, General Knowledge, Reading, and Arithmetic were collected for most of the participants of this study (not all schools used grades to mark their pupils' work). Numbers were transformed to a 5 point-scale, ranging from 1 (very low) to 5 (very high).

Participant-Related Variables

The following variables were used to investigate the influence of participant-related variables on test performance:

Gross National Product (GNP). Participant background information was gathered with individual questionnaires prior to the cognitive data collection. The country of birth of the pupils as well as that of his or her parents was asked. The mean GNP of the three countries (child, mother, and father) was used as GNP measure. Country GNPs were taken from Georgas, Van de Vijver, and Berry (2000).

Ethnic identity. One item of a biographic questionnaire was used as indicator of ethnic identity. It was presented as follows: I feel most like a ... (Dutch, Turkish, Moroccan person). The responses to this open-ended question were scored as follows: 3 = Dutch, 2 = both, and 1 = non Dutch ethnic group.

Number of years spent in the Netherlands. The participant indicated the number of

years spent in the Netherlands by choosing one of the following alternatives: 1 = "less than 2 years", 2 = "2-5 years", 3 = "5-10 years", 4 = "more than 10 years", and 5 = "My whole life".

Preferred language. Three items tapped the language preference when communicating with parents, siblings, and friends. The scoring was as follows: 3 = Dutch, 2 = both, and 1 = ethnic language. The mean of the three responses was used in the analyses.

Frequency of computer usage. This item requested the participant to indicate how often he/she uses a computer. The responses where scored as follows: 4 = every day, 3 = few times a week, 2 = few times a month, and 1 = scarcely.

Computer at home. This item was included to determine whether a computer is present at home. The responses were scored as follows 3 (yes), 2 (computer games), and 1 (no).

Aggregated participant-related variables. The six participant-related variables were factor analyzed for the migrant group. Following Oblimin rotation, two interpretable factors were found, with eigenvalues of 1.86 and 1.31, together explaining 52.9% of the variance (see Table 3). The first factor denoted familiarity with the Dutch society (cultural distance). Participants with a high factor score on this factor are brought up in families born in countries with a high GNP, have lived long in the Netherlands, report to have a Dutch ethnic identity, and prefer to speak Dutch at home. The second factor represented computer experience. Participants with high factor scores have frequently used a computer, and have a computer at home. The two factors show a small, positive correlation of .10.

Table 3. *Structure Matrix of the Two Factors of Eight Participant-Related Variables for the Migrant Samples (Oblimin)*

	Culture familiarity	Computer familiarity
Mean GNP of country of birth of participant and both parents	.82	.10
Ethnic identity[a]	.52	-.20
Years spent in the Netherlands	.67	.13
Preferred language[b]	.62	.24
Frequency computer usage	.06	.82
Computer at home	.13	.79

aHigh score refers to a Dutch identity. bHigh score refers to use of the Dutch language.

Test Characteristics

The following test-related characteristics were examined:

Mean scores. The mean scores of the majority and migrant group are listed in Table 4. The majority group members scored higher on all the tests.

Item difficulty. Item difficulty was operationalized here as the item average, standardized per test (p-values for power tests and reaction times for speed tests).

Aggregated g and c measures. Two test characteristics, g (after Spearman's g) and c (for culture), were derived from analyses on the present data set excluding the first-genera-

tion migrants (see Helms-Lorenz & Van de Vijver, & Poortinga, 2000, for more details). The g and c measures were based on a factor analysis of item and test characteristics. Among the variables loading high on g were an empirical g loading derived from the data as well as two theoretically grounded task complexity measures, derived from the work by Carroll (1993) and Fischer (1980). The variables with the highest loading on the c factor were ratings of the cultural loadings of the tests (by 25 senior psychology students) and the verbal loading of the tests (operationalized as the number of words in a test). The factor scores of the subtests are given in Figure 2 (see chapter 2).

Table 4. *Means of Migrants and Majority Group Members*

	First-generation migrants		Second-generation migrants		Majority	
Test	M	SD	M	SD	M	SD
RAKIT						
Exclusion	37.74	6.34	37.59	6.98	38.91	6.63
Word Meaning	41.00	7.22	41.60	7.52	46.75	6.80
Discs	222.16	95.22	224.06	82.86	162.57	63.39
Learning Names	13.34	4.00	13.85	4.14	17.36	4.08
Hidden Figures	34.00	6.15	33.50	5.55	34.41	5.46
Idea Production	65.27	22.07	60.13	20.78	69.86	20.57
SON-R						
Categories	10.83	4.75	11.75	4.71	13.16	4.71
Analogies	13.50	6.63	15.53	5.79	17.49	6.08
Situations	15.11	5.30	16.38	5.61	18.29	5.27
Mosaics	10.36	3.67	11.45	4.15	13.28	4.21
TAART						
ECT1	1592	487	1554	514	1415	413
ECT2	5006	1637	4765	1478	4517	1555
CITO						
Information	19.07	26.85	34.79	24.20	43.71	28.92
Language	24.47	23.13	39.21	26.46	45.76	28.24
Arithmetic	28.40	25.40	45.24	30.11	46.25	32.49
Grade marks						
Knowledge	2.75	0.98	3.00	0.95	3.54	0.93
Arithmetic	2.79	1.08	2.91	1.08	3.25	1.07
Reading	2.83	1.04	3.06	0.96	3.50	0.94

Note: Higher means point to better performance on all tests except for Discs, and the ECT tasks, which are response times (in seconds and milliseconds, respectively).

Results

Bias Detection

In Table 5 an overview is given of the statistical techniques used to detect construct, item, and method bias.

Table 5. *Bias Detection Techniques used in Analyses*

Type of bias	Detection Technique
Construct Bias	Factor analysis → Tucker's phi and RMSD (test level)
Conditional Item Bias	
Uniform Bias	Anova and logistic regression
Nonuniform Bias	Anova and logistic regression
Unconditional Item Bias	Tucker's phi and RMSD (item level)
Method Bias	Regression analysis → migrant raw scores predicted by participant-related characteristics Correlation between test score and generational status Correlation between test score and average GNP

Construct bias detection. Factor analyses were performed separately for each of the 12 tests in both cultural groups. Tucker's phi (Tucker, 1951), a coefficient of factorial agreement, was calculated for all tests. Values lower than .90 are often taken to point to non-negligible incongruities (Van de Vijver & Leung, 1997). The results are tabulated in Table 6. Discs and Word Meaning showed such incongruities (with values of .87 and .89).

Table 6. *Bias Measures (see text for explanation)*

Test	Tucker's phi	RMSD[a]	Percentage biased items	Percentage items with Uniform bias	Percentage items with Nonuniform bias
RAKIT					
Exclusion	.9225	.1701	9.4	6.3	3.1
Word Meaning	.8913	.2329	46.3	40.7	20.4
Discs	.8729	.2305	22.2	16.7	16.7
Learning Names	.9768	.0928	0	0	0
Hidden Figures	.9414	.1475	10.3	6.9	6.9
Idea Production	.9988	.0412	20	20	20
SON-R					
Categories	.9858	.0840	3.7	0	3.7
Analogies	.9625	.1436	6.1	6.1	3
Situations	.9602	.1400	3.1	3.1	3.1
Mosaics	.9957	.0606	0	0	0
TAART					
ECT1	.9934	.0762	10	10	0
ECT2	.9895	.0919	10	0	10

a RMSD = Root Mean Squared Difference

An additional test-level measure for construct bias, the Root of the Mean Squared Difference (RMSD) was determined for each of the 12 tests. This is the square root of the mean squared difference between the factor loadings of the majority group and the migrants (corrected for differences in eigenvalues of the factors in the two cultures), averaged across all items of a test. Larger values point to more difference between the groups compared. As can be seen in Table 6, Discs (.23), and Word Meaning (.23), and Exclusion (.17) showed the highest RMSD values. It can be concluded that these three tests measure slightly different constructs. The consequences on the test results for the two cultural groups are not clear at this.

Item bias detection. Conditional and unconditional item bias techniques were used. Item-level measures of Tucker's phi and RMSD were used as measures of unconditional item bias. To assess the number of conditionally biased items in each test, the sample was divided into five ability groups. The ability groups were determined for the whole sample (migrants and majority group members combined). For each test a separate analysis was carried out. Analyses of variance (ANOVA) were conducted on the continuous data (Discs, Idea Production, ECT1, and ECT2), while logistic regression analyses were conducted on the remaining, dichotomous data. An item was flagged as uniformly biased when a significant main effect for culture was found, and as nonuniformly biased when the interaction between culture and ability group was significant ($\alpha < .05$).

In Table 6 the percentage of total, uniform and nonuniform biased items per test is listed. The total percentage of biased items is a sum of the two kinds of biased items (correcting for overlap). Some tests showed proportions of biased items close to the .05 level that is the base rate. Three RAKIT tests, Idea Production, Word Meaning, and Discs show higher proportions of biased items compared to the other tests. These results indicate that some of the items of these tests do not measure the same in both groups. In the other tests, the proportions of biased items were not high. That the proportions were somewhat lower than those found in other Dutch educational research (e.g., Kok, 1988; Uiterwijk & Vallen, 1997) may be due to the use of culture-reduced tests in the present study.

Method bias detection. This level of bias occurs if characteristics of the subjects, the tests or their administration are related to test performance when there is no reason for the relationship in terms of the abilities that a test is supposed to measure. The focus is not on the extent to which method bias can explain cross-cultural performance differences, but on the role of the two factors of Table 3, familiarity with the Dutch culture and with computers, within the group of migrants. These constitute noncognitive participant factors (measured only in the migrant group) that should be essentially unrelated to cognitive test performance.

Regression analyses were performed with raw subtest scores as dependent variables (standardized per age group) and non-cognitive participant characteristics, as defined by the two familiarity factors of Table 3, as independent variables. In Table 7 the beta values and the adjusted multiple correlations are presented. Culture familiarity was a

significant predictor of the elementary cognitive tests, the SON-R subtests, two of the six RAKIT subtests (Word Meaning and Idea Production), two of three CITO subtests (Information and Language), and two of the three grade marks (general knowledge and reading). When culture familiarity was a significant predictor, it always yielded a positive contribution. The highest betas were found for the CITO tests. Computer familiarity showed significant contributions to two measures: Idea Production and the grade mark for general knowledge; in both cases a negative coefficient was found. It is remarkable that computer experience was unrelated to performance on the computer-assisted tests (ECT1 and ECT2).

Table 7. *Multiple Regression Analysis with Culture Familiarity and Computer Familiarity as Predictors and Subtest Scores as Dependent Variables*

	Culture familiarity	Computer familiarity	
Test/Subtest	β	β	Adj. R^2
RAKIT (N = 106)			
Exclusion	.00	-.07	-.02
Word Meaning	.27**	-.03	.06*
Discs	-.09	-.05	-.01
Learning Names	.21	-.08	.03
Hidden Figures	-.08	.03	-.01
Idea Production	.01	-.27**	.06*
SON-R (N = 151)			
Categories	.26**	.01	.06**
Analogies	.26**	.03	.05*
Situations	.17*	-.11	.04*
Mosaics	.21*	-.03	.04*
TAART (N = 487)			
ECT1	.12*	-.07	.02**
ECT2	.20***	.01	.03***
CITO (N = 44)			
Information	.48**	-.04	.18**
Language	.37	.21	.16*
Arithmetic	.30	.17	.13
Grade marks (N = 430)			
Knowledge	.12*	-.12*	.02*
Arithmetic	-.03	-.09	.01
Reading	.16**	-.08	.03***

Note. Higher scores point to a higher performance for all tests. *p < .05. **p < .01. ***p < .001.

It can be concluded that the relationships between the factors and performance are small but fairly consistent. In particular the culture familiarity factor was related to performance. The median of the absolute standardized regression coefficient is .19 for the culture factor, but merely .07 for the computer factor.

In a next analysis correlations were computed between test score performance and generation (first or second). As can be seen in Table 8, most tests showed positive, significant correlations, thereby confirming the conclusion of the previous analysis that acquaintance with the Dutch culture, which can be taken to be higher in the second generation, is positively related to school-achievement and mental-performance tests. Higher scores by second-generation migrants have also been reported for adults (Te Nijenhuis, 1997; Van den Berg & Bleichrodt, 2000).

Table 8. *Correlations between Test Performance and Participant Characteristics (Generation Status and Average GNP of the Country of birth of Participants and Parents)*

Test/Subtest	Generation status[a]	GNP
RAKIT		
Exclusion	.21**	.24***
Word Meaning	.34***	.47***
Discs	.22**	.32***
Learning Names	.07	.24***
Hidden Figures	.14	.18**
Idea Production	-.07	.03
SON-R		
Categories	.37***	.32***
Analogies	.38***	.35***
Situations	.36***	.31***
Mosaics	.36***	.38***
TAART		
ECT1	.10**	.24***
ECT2	.22***	.27***
CITO		
Information	.57***	.28**
Language	.44**	.20*
Arithmetic	.50***	.14
Grade marks		
Knowledge	.10	.24***
Arithmetic	.08*	.11***
Reading	.16***	.23***

Note. Higher scores point to a higher performance for all tests. a1 = first generation; 2 = second generation. *$p < .05$. **$p < .01$. ***$p < .001$.

Finally, the analysis was extended to the group of mainstream pupils. For the combined samples of participants correlations were computed between test performance and the average GNP of the pupil's country of birth and of his or her parents (average of the three values). For most measures the correlations were positive and highly significant. So, both generation status and the GNP measure pointed in the same direction.

Antecedents of Item Bias

The measures of item bias, root mean squared difference, Tucker's phi, uniform bias, and nonuniform bias (the latter two are dichotomous bias indicators derived from logistic regression and analysis of variance as described above) were factor analyzed. Two factors with eigenvalues of 1.78 and 1.11 were extracted, explaining 72.2% of the variance. The loadings (after Varimax rotation) are presented in Table 9. The first factor combined the first two, factor-analytically derived, bias statistics, while the latter two statistics showed high loadings on the second factor. The first factor was labeled unconditional bias (as the statistics are based on analyses that do not take score level into account) while the second factor represented conditional bias.

Table 9. *Loadings of the Two Factors Derived from a Factor Analysis on Four Item-Bias Measures (Varimax-Rotated)*

	Unconditional bias	Conditional bias
Root Mean Squared Difference	-.85	.08
Tucker's phi	.84	-.12
Uniform bias	.10	.84
Nonuniform bias	.09	.84

The influence of item and test characteristics (item difficulty, internal consistency, g factor, and c factor) on the two factors was examined in a regression analysis. Item difficulty was included as it has been found to predict item bias (Linn, 1993). Scores of the unconditional bias factor could be relatively well predicted (R^2 = .29, p < .001)(see Table 10). More difficult items (β = -0.16) and tests with a higher internal consistency (β = -0.33), higher cognitive complexity (β = -0.22), and more cultural loading (β = 0.25) were found to show less item bias (all ps < .001). However, the prediction of conditional bias was less successful. Item difficulty was the only significant predictor; its effect was (again) negative, β = -0.13, p < .05. The multiple correlation was low (R^2 = .01) and nonsignificant. The failure to identify antecedents of conditional bias statistics replicates American studies of antecedents of item bias (e.g., Bond, 1993; Linn, 1993; Scheuneman, 1987).

Table 10. *Regression Analysis of Item Bias Factors (Standardized Regression Coefficients)*

	Dependent Variable	
Predictor	Unconditional Bias	Conditional Bias
Item difficulty[a]	-.16**	-.13*
Internal consistency	-.33***	.00
g factor	-.22***	.08
c factor	.25***	-.06
Adj. R^2	.29***	.01

aHigher score refers to more difficult item. *p < .05. **p < .01. ***p < .001.

The remarkable difference in the predictability of unconditional and conditional item bias statistics deserves closer scrutiny. When the RMDS statistic was computed for each item, the items with a nonsignificant uniform bias showed a significantly lower mean on the RMSD than did the items with significant uniform bias (means of 0.11 and 0.16, respectively; $t(271) = -2.41$, $p < .05$). The same was found for the nonuniformly biased items (means of 0.10 and 0.17, $t(271) = -2.81$, $p < .01$). Correlations between both types of statistics were significant but low (uniform: .14; nonuniform: .17, for both $p < .05$). Different explanations could be envisaged for this low correspondence. The first could be discriminatory power: conditional analyses use more fine-grained analyses and hence, may pick up more bias. The present data do not support this interpretation. Not all items with a high RMSD are flagged as biased by conditional methods.

Alternatively, conditional and unconditional bias statistics may be susceptible to different sources of distortion. More specifically, conditional techniques identify sources of error that may go unnoticed when using unconditional bias statistics, such as a slightly different meaning of an item for low and high scorers in one cultural group and floor or ceiling effects in groups with extreme score in one cultural group. The latter are mere method artifacts that may be related only to item difficulty. Also, more subtle bias mechanisms such as shifts in meaning with score level are not identified by simple item- or test characteristics. The diversity of antecedent factors may make understandable the poor predictability of conditional bias statistics.

Discussion

The first research question addressed the role of construct, method, and item bias in the performance on culture-reduced tests in a multicultural population. We found that construct- and item bias, were present in a small number of tests. Method bias, however, is clearly present in culture-reduced tests and influences the performance of migrant pupils, significantly. All three bias detection measures flagged that something is amiss. The three forms of bias tell the same message, although only method bias makes what the size of the impact is on the test scores.

Measures used in this study to assess construct bias were Tucker's phi and RMSD. Ten of the twelve tests showed acceptable levels of structural equivalence. Thus, for these tests no construct bias was found. For the remaining two tests, Word Meaning and Discs, construct bias was found. In other words, these tests are likely to assess different aspects of psychological functioning among migrant children and children from Dutch born parents.

Item bias detection was performed using logistic regression and ANOVA. The overall results revealed that three tests had biased items. This indicated that the items cannot be assumed to have been sampled from corresponding ability domains.

This was demonstrated by the analysis of method bias. Evidence for the presence of method bias was derived from different sources. First, culture familiarity was a rela-

tively good predictor of migrant pupils' performance on more crystallized achievement measures (CITO tests), but was relatively unsuccessful in predicting performance on less crystallized measures of ability. Second, generation status was found to predict migrant pupils performance on various tasks, notably the more crystallized achievement measures. Third, if data of both natives and migrants were taken into consideration, the average GNP of the country of birth of the pupil and his or her parents showed a significant relationship with performance; again, the correlations for the more crystallized measures were stronger. The relevance of culture familiarity on performance and the moderating role of the degree of crystallization of the tasks provide strong evidence for the presence of method bias.

The construct and item bias results are in line with the overall research results in this field. If our conclusions were to be based on the results of these studies only, we would have concluded that bias does not play a major role in the mental test performance of migrant children. The picture changes considerably when method bias is examined. Noncognitive participant-related factors were significant predictors of migrant test performance. These results are in line with Helms-Lorenz and Van de Vijver (2000) who found that the $\underline{c}$ factor was at least as important as cognitive complexity in the explanation of performance differences of the majority-group and migrants.

Two implications emerge from the present study. First, the emphasis in the literature on structural equivalence and item bias may lead to an underestimation of the influence of method factors. The more widely known and perhaps better methods of analysis for construct and item bias than for method bias should not be interpreted as a sign of the irrelevance of the latter. Quite the contrary, from a cross-cultural perspective it is hard to understand why the study of method-related factors has been neglected. Focusing on a specific, single form of bias can yield a distorted picture of the validity of intergroup comparisons. It is only through an inclusive and balanced treatment of different sources of bias that we gain insight in the nature of observed cross-cultural similarities and differences.

Second, the generalizability of the results should be addressed. Will similar results hold elsewhere (e.g., in the USA)? Some characteristics of the migrant children studied are specific to Western Europe, such as the high prevalence of Mediterranean, Islamic groups. Other characteristics, however, are common to various migrant school children, such as the relatively low SES of the parents. The samples studied here have the typical language problems and underprivileged position in society shared by many migrated groups. More pronounced method bias effects are likely to be found among first-generation migrants. Analogously, more pronounced bias effects might be found when using instruments that have not been designed for usage in multicultural groups. It is unlikely that the present results can be dismissed as mere sample and test peculiarities.

References

Arvey, R. D., Strickland, W., Drauden, G., & Martin, C. (1990). Motivational components of test taking. *Personnel Psychology, 43*, 695-715.

Bleichrodt, N., Drenth, P. J. D., Zaal, J. N., & Resing, W. C. M. (1984). *Revisie Amsterdamse Kinder Intelligentie Test*. Lisse: Swets & Zeitlinger Bleichrodt, N., & Van de Vijver F. J. R. (Eds.) (2000). *Diagnostiek bij allochtonen.* Lisse, the Netherlands: Swets & Zeitlinger.

Bond, L. (1993). Comments on the O'Neill and McPeek's paper. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 277-279). Hillsdale, NJ: Erlbaum.

Burton, E., & and Burton, N. (1993). The effect of item screening on test scores and test characteristics. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 321-335). Hillsdale, NJ: Erlbaum.

Carroll, J. B. (1993). *Human cognitive abilities. A survey of factor-analytic studies.* Cambridge: Cambridge University Press.

Chan, D., & Schmitt, N. (1997). Video-based versus paper-and-pencil method of assessment in situational judgement tests: subgroup differences in test performance and face validity perceptions. *Journal of Applied Psychology, 82*, 143-159.

Chan, W., Ho, R. M., Leung, K., Cha, D. K-S., & Yung, Y-F. (1999). An alternative method for evaluating congruence coefficients with Procrustes rotation: A bootstrap procedure. *Psychological Methods, 4,* 378-402.

Deregowski, J. B., & Serpell, R. (1971). Performance on a sorting task: A cross-cultural experiment. *International Journal of Psychology, 6,* 273-281.

Evers, A., Van Vliet-Mulder, & Ter Laak, J. (1992). *Documentatie van Tests en Testresearch in Nederland.* Assen, Van Gorcum.

Eysenck, H. J., Barrett, P., & Eysenck, S. B. (1985). Indices of factor comparison for homologous and non-homologous personality scales in 24 different countries. *Personality and Individual Differences, 6,* 503-504.

Fischer, K. W. (1980). A theory of cognitive development: the control and construction of hierarchies of skills. Psychological Review, 87, 477-531.

Foorman, B. R., Yoshida, H., Swank, P. R., & Garson, J. (1989). Japanese and American children's styles of processing figural matrices. *Journal of Cross-Cultural Psychology, 20,* 263-295.

Georgas, J., Van de Vijver, F. J. R., & Berry, J. W. (2000). *Ecosocial indices and psychological variables in cross-cultural research* (under review).

Hambleton, R. K., & Van der Linden, W. J. (Eds.) (1997). Handbook of Modern Item Response Theory. New York: Springer Verlag.

Helms, J. E. (1992). Why is there no study of cultural equivalence in standardized cognitive ability testing? *American Psychologist, 47*, 1083-1101.

Helms-Lorenz, M., & Van de Vijver F. J. R. (1995). Cognitive assessment in education in a multi-cultural society. *European Journal of Psychological Assessment, 11,* 158-169.

Helms-Lorenz, M., Van de Vijver, F. J. R., & Poortinga, Y. H. (2000). Cross-Cultural Differences in Cognitive Performance and Spearman's Hypothesis: G or C? Tilburg, the Netherlands: Tilburg University. (under review)

Holland, P. W., & Wainer, H. (Eds.) (1993). *Differential Item Functioning.* Hillsdale, NJ: Erlbaum.

Hui, C. H., & Triandis, H. C. (1989). Effects of culture and response format on extreme response style. *Journal of Cross-Cultural Psychology, 20,* 296-309.

Irvine, S. H. (1979). The place of factor analysis in cross-cultural methodology and its contribution to cognitive theory. In L. Eckensberger, W. Lonner, & Y. P. Poortinga (Eds.), *Cross-cultural contributions to psychology.* Lisse, the Netherlands: Swets & Zeitlinger.

Jensen, A. R. (1980). *Bias in mental testing.* New York: Free Press.

Jensen, A. R. (1985a). The nature of Black–White differences on various psychometric tests: Spearman's hypothesis. *Behavioral and Brain Sciences, 8,* 193-263.

Kok, F. (1988). *Vraagpartijdigheid. Methodologische verkenningen* [Item bias. Methodological explorations]. Amsterdam: University of Amsterdam (Thesis).

Laros, J. A., & Tellegen, P. J. (1991). *Construction and validation of the SON-R 5 ½- 17, the Snijders-Oomen non-verbal intelligence test.* Groningen, the Netherlands: Wolters-Noordhoff.

Linn, R. L. (1993). The use of differential item functioning statistics: A discussion of current practice and future implications. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 349-364). Hillsdale, NJ: Erlbaum.

Martens, E. P., & Veenman, J. (1999). De sociaal-economische positie van etnische minderheden [The social-economical position of ethnic minorities]. In H. M. A. G. Smeets, E. P. Martens, & J. Veenman (Eds.), *Jaarboek minderheden* (pp. 107-138). Houten, the Netherlands: Bohn Stafleu Van Loghum.

McCrae, R. R., & Costa, P. T. (1997). Personality trait structure as a human universal. *American Psychologist, 52,* 509-516.

Mellenbergh, G. J. (1982). Contingency table models for assessing item bias. *Journal of Educational Statistics, 7,* 105-118.

Mercer, J. R. (1984). What is a racially and culturally nondiscriminatory test? In C. R. Reynolds & R. T. Brown (Eds.), *Perspectives on bias in mental testing* (pp. 293-356). New York: Plenum.

Rogers, W. T., & Yang, P. (1996). Test-wiseness: Its nature and application. *European Journal of Psychological Assessment, 12,* 247-259.

Scheuneman, J. (1979). A method of assessing bias in test items. *Journal of Educational Measurement, 16,* 143-152.

Scheuneman, J. 1987). An experimental, exploratory study of causes of bias in test items. *Journal of Educational Measurement, 24,* 97-118.

Schmitt, A. P. (1988). Language and cultural characteristics that explain differential item functioning for Hispanic examinees on the scholastic aptitude test. *Journal of Educational Measurement, 25,* 1-13.

Serpell, R. (1979). How specific are perceptual skills? *British Journal of Psychology, 70,* 365-380.

Shepard, L. A., Camilli, G. & Averill, M. (1981). Comparison of procedures for detecting test-item bias with both internal and external ability criteria. *Journal of Educational Statistics, 6,* 317-375.

Smith, J. D., & Caplan, J. (1988). Cultural differences in cognitive style development.

Developmental Psychology, 24, 46-52.

Te Nijenhuis, J. (1997). Comparability of test scores for immigrants and majority group members in the Netherlands. Amsterdam: Free University.

Uiterwijk, H., & Vallen, T. (1997). Onderzoek naar bias voor allochtone leerlingen in de Cito-Eindtoets Basisonderwijs. *Pedagogische Studiën, 74,* 21-32.

Van de Vijver, F. J. R. (1997). Meta-analysis of cross-cultural comparisons of cognitive test performance. *Journal of Cross-Cultural Psychology, 28,* 678-709.

Tucker, L. R. (1951). *A method for synthesis of factor analysis studies* (Personnel Research Section Report No. 984). Washington, DC: Department of the Army.

Van de Vijver, F. J. R., Daal, M., & Van Zonneveld, R. (1986). The trainability of formal thinking: A cross-cultural comparison. *International Journal of Psychology, 21,* 589-615.

Van de Vijver, F. J. R., & Leung, K. (1997). *Methods and data analysis for cross-cultural research.* Thousand Oaks, CA: Sage.

Van de Vijver, F. J. R., & Willemse, G. R. (1991). Are reaction time tasks better suited for ethnic minorities than paper-and-pencil tests? In N. Bleichrodt & P. J. D. Drenth (Eds.), *Contemporary issues in cross-cultural psychology* (pp. 450-464). Lisse, the Netherlands: Swets & Zeitlinger.

Van den Berg, R. H., & Bleichrodt, N. (2000). Het meten van cognitieve vaardigheden bij allochtone volwassenen [The measurement of cognitive skills of migrant adults]. In N. Bleichrodt & F. J. R. Van de Vijver (Eds.), Diagnostiek bij allochtonen (pp. 104-125). Lisse, the Netherlands: Swets.

Verhoeven, L. (2000). Detectie van taalachterstand [Detection of language deficiency]. In N. Bleichrodt & F. J. R. Van de Vijver (Eds.), *Diagnostiek bij allochtonen* (pp. 180-204). Lisse, the Netherlands: Swets & Zeitlinger.

Epilogue

The development of the cognitive computerized reaction time test, TAART, has resulted in a test battery that forms an alternative and an extension to the already available culture-reduced tests. The principles described in the first chapter were applied in the test design in order to minimize potentially biasing subject- and instrument-related factors. However, this does not yet mean that cultural loadings were fully eliminated. The mean cultural loading ratings of TAART by senior psychology students (Table 2, Chapter 2) fall within the values found for RAKIT and SONR tests, be it at the lower end of the range. Therefore, we conclude that TAART is culture-reduced but certainly not acultural, culture-free, or culture-fair. This is in line with all previous attempts in the past to develop culture-free and fair tests, which led Frijda and Jahoda already in 1969 to the conclusion that there is no such thing as a culture-free or fair test.

Compared to currently available culture-reduced tests in the Netherlands, TAART has a number of practical advantages. The test administration is simplified to a large extent because of its group administration, automated test presentation and response registration. These aspects minimize the likelihood of errors made by the tester.

The results of the theoretical part of this study can be summarized as follows. The unraveling of the first factor loadings (*g*-factor) described in Chapter 2, disclosed the confounding of task complexity with other test characteristics. Our choice to use culture-reduced tests minimized the influence of bias factors. The evidence indicating that even in our culture-reduced tests *g* (as measured) is more than *g* should be (i.e., a pure measure of cognitive complexity), strongly suggests even greater confounding in culturally more loaded tests.

An aggregated culture factor (the *c*-factor) was found to be present in all the culture-reduced tests, the TAART tasks, as well as the tests taken from other existing batteries. There was evidence that the *c*-factor was at least as important as cognitive complexity in the explanation of performance differences of majority group members and migrant children in The Netherlands.

In the third chapter the role of bias on test results was examined. Evidence was found for structural equivalence for most of the tests; but for two of the twelve tests there was evidence of structural inequivalence. Results indicated fairly extensive item bias although a direct reflection on the performance could not be demonstrated. Evidence for the presence of method bias was derived from different sources. First, culture familiarity was a relatively good predictor of migrant pupils' performance on more crystallized achievement measures (CITO tests), but was relatively unsuccessful in predicting performance on less crystallized measures of ability. Second, generation status was found to predict migrant performance on various tasks, notably the more crystallized achievement measures. Third, if data of both natives and migrants were taken into consideration, the average GNP of the country of birth of the pupil and his or her par-

ents showed a significant relationship with performance; again, the correlations for the more crystallized measures were stronger. The relevance of culture familiarity on performance and the moderating role of the degree of crystallization of the tasks provide strong evidence for the presence of method bias. Furthermore, the relationship of test characteristics (item difficulty, internal consistency, g factor and c factor) with item bias provided an interesting extension of existing literature in which typically item bias of a single test is predicted on the basis of item parameters of a test, usually without much success (e.g., Scheuneman, 1987).

A combination of the results of Chapters 2 and 3 paves the way to the conclusion that most likely a substantial part of the observed performance differences on cognitive tests can be attributed to cultural bias in the instruments used in the multicultural context of the Netherlands.

The present findings demonstrate that method/test or cultural bias is not to be underestimated. In a review of bias research in The Netherlands, Te Nijenhuis and Van de Vijver (2000) conclude that the effects of test bias are not strong. Their conclusion is based on (internal) item bias studies (factorial studies and item bias detection), and (external) studies of prediction bias with criteria that may well share important components of bias with the tests. The present study clearly demonstrates that an analysis of bias should proceed beyond the level of separate items and should also examine more global validity threats (method bias). Internal bias studies as reported by The Nijenhuis and Van de Vijver are not sufficient to rule out other types of bias than item bias.

Moreover, in many of the cited studies, effect sizes were not reported, thus not permitting further analysis of the patterning of cross-cultural score differences (e.g., Resing, Bleichrodt, & Drenth, 1986; Uiterwijk & Vallen, 1997). Similarly, the positive evidence quoted by Te Nijenhuis (1997) supporting Spearman's hypothesis does not reject the cultural bias hypothesis since the possible confounding of the *g*-factor with the *c*-factor was not considered (see Chapter 2). In our opinion Jensen (1985) fails to appreciate the possibility of a confounded *g* factor. Positive evidence to support Spearman's Hypothesis does not rule out the influence of other (even more potent) influences on cognitive performance. Convergent validation (aiming at relationships between variables that are expected to be related) is providing only part of the evidence. There is also a need to establish disciminant validity (aiming at finding absence of relationships when theoretically no relationship should be found). Discriminant evidence is a powerful tool to be used to test the monopoly of a theory. In our opinion the discriminant validity of Spearman's Hypothesis had never been seriously considered until now.

Equivalence is a property of a specific cross-cultural comparison. It is a function of characteristics of an instrument and of the cultural groups involved. Therefore, future studies should determine the generalizability of the present findings to new instruments and cultural populations.

Implications

Many of the practical and theoretical implications of this study have already been alluded to. We like to point explicitly to the following.

Practical: The research performed in the past and based on the results of this project, we are confident to conclude that all instruments are culturally loaded, be it to different extents. The measurement of 'pure g' is a folly. If the intention of test administration is to determine a migrant pupil's intelligence (cognitive ability) without accounting for knowledge of the language and culture of the majority group, a culture-reduced test should be preferred. Even if a culture-reduced test is not free of cultural loading, it will offer the best attainable estimation of cognitive ability. The assessment of the cognitive abilities of pupils entering the school system, who have only come to live in the Netherlands, recently should be done with culture-reduced tests only. When however, future behavior is to be predicted where knowledge of the language and culture of the majority group is relevant, then culturally more loaded instruments can be used in conjunction with culture-reduced tests, as these may well show stronger correlations with the criterion behavior (school success, or success on the job). The test results will reveal the candidate's cognitive potential as well as his/her progress in "doing things the majority group way". Ultimately this could lead to better substantiated decisions and interventions: if a candidate shows high cognitive potential, but low knowledge of the language and culture of the majority group, then his/her cognitive potential would be underestimated if only culturally loaded tests were to be considered.

Theoretical: Our results have shown that both the *g*-factor as well as the *c*-factor derived from tests is responsible for performance differences found between migrants and majority groups. We have to critically examine assessment procedures in multicultural settings. Cultural loadings are more influential and difficult to reduce than commonly acknowledged.

References

Jensen, A. R. (1985). The nature of Black–White difference on various psychometric tests: Spearman's hypothesis. *Behavioral and Brain Sciences, 8,* 193-263.

Resing, W. C. M., Bleichrodt, N., & Drenth, P. J. D. (1986). Het gebruik van de RAKIT bij allochtoon etnische groepen. *Nederlands Tijdschrift voor de Psychologie, 41,* 179-188.

Scheuneman, J. D. (1987). An experimental, exploratory study of causes of bias in test items. *Journal of Educational Measurement, 24,* 97-118.

Te Nijenhuis, J. (1997). Comparability of test scores for immigrants and majority group members in the Netherlands. Amsterdam: Free University.

Te Nijenhuis, J., & Van de Vijver, F. J. R. (2000). Onderzoek naar partijdigheid. In F. J. R. Van de Vijver & N. Bleichrodt (Eds.), Diagnostiek bij allochtonen (pp. 49-75). Lisse, the Netherlands: Swets & Zeitlinger.

Uiterwijk, H., & Vallen, T. (1997). Onderzoek naar bias voor allochtone leerlingen in de Cito-Eindtoets Basisonderwijs. *Pedagogische Studiën, 74,* 21-32.

Vernon, P. A. (1969). *Intelligence and cultural environment. London: Methuen.*

Summary

The population in the Netherlands has changed from a fairly homogeneous cultural group to a heterogeneous one over the last four decades. This diverse multicultural population has formed a new challenge for the educational system. Commonly used cognitive and educational tests are to be scrutinized for use in multicultural groups.

Examinations of the suitability of ability tests in multicultural applications was stimulated by repeated findings that subject- and instrument-related factors negatively influence cognitive performances of minority group members. In Chapter 1 this point is elaborated; it supplies guidelines to avoid typical pitfalls in multicultural assessment. Various factors are discussed that can challenge the equivalence (and hence, the comparability) of the test scores obtained in these groups, such as intergroup differences in verbal skills, in cultural values and norms, and in test-wiseness. Commonly applied remedies to enhance the suitability of cognitive tests are discussed: adaptation of existing tests, the use of different norms, statistical and linguistic procedures to correct for item bias, and the development of new tests.

A computer-assisted elementary cognitive test battery called TAART (an acronymn for Tilburgse Allochtonen en Autochtonen Reactie Tijd Test) was further developed for this project. In developing this instrument the most important objective was to reduce the influence of potentially biasing subject-related factors on test performance, such as verbal skills, cultural knowledge, and test-wiseness. The test is virtually non-verbal. It runs on IBM-compatible computers and uses the mouse as response device. The whole battery consists of nine subtests; results of the only two subtests that were administered to all age groups are reported here. Geometric figures are used in the items. In the first task (ECT1) five figures are shown, consisting of two pairs of identical stimuli and an "odd one out." The participant has to identify the latter. The second task (ECT2) involves "complementary figures." Complementary figures form exactly one black square when they are "added" (combined). Each ECT2 item consisted of two pairs of complementary figures and an "odd one out." The latter had to be identified by the pupil.

Both ECT1 and ECT2 consist of two series of ten items each, with a short break in between. When an item is presented on the screen, the mouse is located in the center of the screen in the "mouse box." This mouse box is surrounded by five squares, all at equal distance from the mouse box in a circular arrangement. The reaction time (used as performance measure) is defined as the time elapsed between stimulus onset and the moment the pupil moves the mouse outside the borders of the mouse box. In order to ensure that the pupil identifies the target figure before starting to move the mouse, the contents of the squares become gray and only the borders remain visible once the mouse leaves the mouse box. Pupils were instructed to respond fast without making any errors.

Both tests have four practice items. The computer gives feedback about correctness of responses (a face appears on the screen that is either happy or sad). The practice items

are administered again if one or more incorrect responses are given. The actual testing starts when all four exercise items have been solved correctly.

Empirical studies often show that migrant pupils score consistently lower on cognitive tests than native pupils in the Netherlands. The central topic of this project is to address two explanations for cross-cultural performance differences: first, these performance differences can be real and second they can be the product of subject- and/or instrument related factors.

Chapter 2 deals with the investigation of the first explanation; more specifically Spearman's Hypothesis (SH) is tested. SH was put forward by Jensen (1985), and is based on Spearman's observation in 1927 that performance differences between cultural groups increase as tasks become more complex. Jensen operationalized task complexity in factor analytic terms, as being the factor loadings on the first factor (called the test's *g loading*). Now, SH states that performance differences between cultural groups on cognitive tests increase with *g loading*.

In this project the adequacy of the operationalization of complexity in terms of *g* loading to test SH is examined. In our analysis an attempt is made to decompose *g* in verbal–cultural aspects and cognitive complexity. The relative contribution of complexity and verbal–cultural factors to observed cross-cultural performance differences is compared. A sample of 1228 primary school children, age 6 to 12 years, were selected from different regions in the Netherlands (the six- and seven-year old children were combined in the analyses). The sample consisted of Dutch majority-group members ($\underline{n}$ = 747), and a group of second-generation migrants ($\underline{n}$ = 474). In both cultural groups half were boys and half were girls. The majority of the participants were tested in urban regions where migrants mainly reside.

No first-generation migrants were involved in the study. These tend to have a lower level of knowledge of the Dutch language and culture than second-generation migrants. The inclusion of first-generation migrants would have boosted the influence of verbal–cultural aspects on test performance. Restricting the study to second-generation children assured that all children studied had followed a known (and across cultural groups equal) number of years of Dutch education, and had sufficient command of Dutch for the test administration.

Yet, there is evidence that substantial differences in knowledge of the Dutch lexicon between majority-group pupils and migrant pupils linger on throughout the primary school period, even for second-generation children (Verhoeven, 2000). A fairly large number of culture-reduced tests were administered to this multicultural sample of school children in The Netherlands. Three g measures were used. The first defined as by Jensen, the second was derived from the Skill Theory (Fischer, 1980), and the third measure was derived from Carroll's (1993) hierarchical model of the structure of cognitive abilities. Senior psychology students rated the cultural loadings of all the items and the tests. The verbal loading of the tests was determined by counting words. These variables were factor analyzed yielding two aggregated g and c factors. These fac-

tors were used to predict migrant-majority performance differences. The findings of this study suggest that performance differences between majority-group members and migrant pupils are better predicted by a cultural factor (c) than by g.

On the basis of research performed in the past and based on the results of this study, we are confident to conclude that all instruments are culturally loaded, be it to different extents. The measurement of 'pure g' is a folly. If the intention of test administration is to determine a migrant pupil's intelligence (cognitive ability) without accounting for knowledge of the language and culture of the majority group, a culture-reduced test should be preferred. Even if a culture-reduced test is not free of cultural loading, it will offer the best attainable estimation of cognitive ability. The assessment of the cognitive abilities of pupils entering the school system, who have come to live in the Netherlands, recently should be done with culture-reduced tests only. When however, future behavior is to be predicted where knowledge of the language and culture of the majority group is relevant, culturally more loaded instruments can be used. The reason for this is that tests may well show stronger correlations with the criterion behavior (school success, or success on the job) compared to culturally reduced tests. It is advisable to use these tests in conjunction with culture-reduced tests, because the test results will reveal the candidate's cognitive potential as well as his/her progress in "doing things the majority group way". Ultimately this could lead to better substantiated decisions and interventions: if a candidate shows high cognitive potential, but low knowledge of the language and culture of the majority group, then his/her cognitive potential would be underestimated if culturally loaded tests were to be considered only.

The second aim of this project, presented in Chapter 3, was to detect construct, method, and item bias in different cognitive and educational measures that were designed for use in the Netherlands and to identify some of the antecedents of bias. Twelve culture-reduced subtests derived from two standardized intelligence batteries as well as two TAART tasks were administered to the sample. Exploratory factor analytic solutions of subtest scores obtained in both cultural groups were compared to assess construct bias. The analysis of method bias, based on migrant data, estimated the influence of non-cognitive participant characteristics (acquaintance with the Dutch culture and with computers) on test performance. Item bias was assessed using both logistic regression and ANOVA. Effects of construct and item bias could be identified in some tests, but method bias was found in almost all the tests used.

Evidence for the presence of method bias was derived from different sources. First, culture familiarity was a relatively good predictor of migrant pupils' performance on more crystallized achievement measures (CITO tests), but was relatively unsuccessful in predicting performance on less crystallized measures of ability. The median of the absolute standardized regression coefficient was .19 for cultural familiarity. Second, generation status was found to predict migrant pupils performance on various tasks, notably the more crystallized achievement measures. Third, if data of both natives and migrants were taken into consideration, the average GNP of the country of birth of the

pupil and his or her parents showed a significant relationship with performance; again, the correlations for the more crystallized measures were stronger. The relevance of culture familiarity on performance and the moderating role of the degree of crystallization of the tasks provide strong evidence for the presence of method bias.

The construct and item bias results are in line with the overall research findings in this field. If our conclusions would have been based on the results of these studies only, we might have concluded that bias does not play a major role in the mental test performance of migrant children. The picture changes when method bias is examined. We could actually demonstrate that noncognitive participant-related factors were significant predictors of migrant performance. These results are in line with Chapter 2 where the $\underline{c}$ factor was found to be at least as important as cognitive complexity in the explanation of performance differences of the majority-group and migrants. A combination of the results of Chapters 2 and 3 paves the way to the conclusion that most likely a substantial part of the observed performance differences on cognitive tests can be attributed to cultural factors of the instruments used in the multicultural context of the Netherlands.

Finally, antecedents were analyzed for conditional and unconditional measures of item bias. Test characteristics (item difficulty, internal consistency, $\underline{g}$ factor, and $\underline{c}$ factor) were used to predict unconditional and conditional bias. Scores on the unconditional bias factor could be relatively well predicted ($\underline{R}^2 = .29$, $\underline{p} < .001$). Scores on the conditional bias factor could not be predicted successfully. Conditional and unconditional bias statistics may be susceptible to different sources of distortion. More specifically, conditional techniques identify sources of error that may go unnoticed when using unconditional bias statistics. Simple item- or test characteristics might not be able to identify more subtle bias mechanisms such as shifts in meaning with score level. This may explain the poor predictability of conditional bias statistics.

In summary, two findings emerge from the last study. First, the emphasis in the literature on structural equivalence and item bias may lead to an underestimation of the influence of cultural bias. The more widely known and perhaps better methods of analysis for construct and item bias should not be interpreted as a sign of the irrelevance of method bias. Quite the contrary, from a cross-cultural perspective it is hard to understand why the study of method-related factors has been neglected so much. Ignoring one important form of bias can yield a distorted picture of the validity of intergroup comparisons. It is only through an inclusive and balanced treatment of different sources of bias that we gain insight in the nature of observed cross-cultural similarities and differences.

Samenvatting

De Nederlandse samenleving krijgt steeds meer een duurzaam multicultureel karakter. Dit heeft implicaties voor onder andere het onderwijs dat al een aantal jaren te maken met een instroom van leerlingen met een allochtone culturele achtergrond. Dit vergt speciale aandacht van leerkrachten en onderwijsbegeleiders. Zo rijst de vraag naar de bruikbaarheid van reguliere psychologische tests om cognitieve vaardigheden van met name allochtone leerlingen vast te stellen. Veel tests verwijzen expliciet of impliciet naar de Nederlandse taal en cultuur, ook als deze kennis niet zelf het onderwerp van de test vormt.

Onderzoek naar de bruikbaarheid van vaardigheidstoetsen bij multiculturele groepen werd gestimuleerd door de herhaalde bevinding dat persoons- en instrumentgerelateerde factoren de cognitieve prestaties van leden van minderheidsgroepen negatief beïnvloeden. In hoofdstuk 1 wordt dit punt nader uitgewerkt; het verschaft richtlijnen om typische valkuilen te vermijden. Een verscheidenheid aan factoren wordt besproken die de equivalentie van scores in gevaar kan brengen (waardoor de vergelijkbaarheid van scores omlaag gaat), zoals groepsverschillen in verbale vaardigheden, verschillen in culturele waarden en normen, en test-wiseness. In dit hoofdstuk wordt ook een aantal remedies besproken die de bruikbaarheid van cognitieve toetsen verhoogt: test adaptatie, het gebruik van aparte normtabellen per groep, statistische en linguïstische procedures om voor bias te corrigeren, en de ontwikkeling van nieuwe tests.

Voor dit project is er een computer ondersteunde test batterij genaamd TAART (Tilburgse Allochtonen en Autochtonen Reactie Tijd Test) ontwikkeld. Bij de ontwikkeling van deze test was de belangrijkste doestelling het reduceren van de potentiële invloed van persoonsgerelateerde factoren op test prestaties, zoals verbale vaardigheden, culturele kennis en test-wiseness. De test is grotendeels non-verbaal. Het programma werkt op IBM-compatible computers en de muis fungeert als respons apparaat. De testbatterij bevat negen subtests; twee subtests zijn bij alle proefpersonen afgenomen en de resultaten van deze tests worden hier gepresenteerd. Geometrische figuren zijn gebruikt in de items. De eerste taak (ECT1) bestaat uit twee paren identieke figuren en een alleenstaand figuur. Van de participant wordt gevraagd het figuur zonder "vriendje" te identificeren. In de tweede taak (ECT2) verschijnen twee paren complementaire figuren op het scherm, terwijl van een van de figuren geen complement aanwezig is. Twee figuren zijn complementair als deze bij mentale samenvoeging precies een gevuld vierkant vormen. De taak is wederom om het figuur zonder "vriendje" te identificeren.

Beide de ECT1 en ECT2 taken bestaan uit twee series van tien items elk, met een korte pauze tussen de series. Bij presentatie op het scherm, is de muis in het middelpunt van het scherm gepositioneerd in het "muizenhok". Het muizenhok wordt omgeven door de 5 figuren die concentrisch eromheen gerangschikt zijn. De reactie tijd (dat als prestatiemaat wordt gebruikt) wordt gedefinieerd als de verstreken tijd tussen

stimulus aanbieding en het moment dat het subject de muis over de grens van het muizenhok beweegt. Om er van zeker te zijn dat het subject het doelwit figuur mentaal gekozen heeft alvorens de muis te bewegen, wordt de inhoud van de figuren grijs met slechts de grenzen ervan zichtbaar zodra de muis het muizenhok verlaat. De subjecten worden geïnstrueerd zo snel mogelijk te reageren zonder om fouten te maken.

Beide subtests hebben vier oefen items. De computer geeft feedback over de juistheid van de respons. (Een gezicht verschijnt op het scherm dat óf vrolijk óf verdrietig is) De oefen items worden opnieuw aangeboden indien één of meer fouten bij het oefenen zijn geconstateerd. De eigenlijke test begint pas als alle vier de oefen items foutloos zijn gemaakt.

Empirische studies wijzen herhaaldelijk uit dat allochtone leerlingen op consistente wijze lager scoren op cognitieve tests dan autochtone leerlingen. In Nederland, maar ook elders. Het centrale onderwerp van dit project is om twee verklaringen voor deze cross culturele prestatieverschillen te bestuderen: ten eerste, deze prestatieverschillen zijn echt en ten tweede zij zijn het product van persoons- en/of instrument gerelateerde factoren.

Hoofdstuk 2 behandelt de eerste verklaring: meer specifiek wordt Spearman's Hypothesis (SH) getoetst. SH is geformuleerd door Jensen (1985), en is gebaseerd op Spearman's observatie in 1927 dat prestatieverschillen tussen culturele groepen toenemen naarmate taakcomplexiteit toeneemt. Jensen operationaliseerde taakcomplexiteit in factoranalytische termen, zijnde de factorladingen op de eerste factor (de g lading van de test genoemd). Jensen definieerde SH als volgd: de prestatieverschillen tussen culturele groepen op cognitieve tests nemen toe naarmate de g lading toeneemt.

In dit project wordt onderzocht of de operationalisatie van taakcomplexiteit in termen van g lading adequaat is om SH te toetsen. In onze analyses wordt een poging gedaan om g te ontrafelen in verbaalculturele aspecten en cognitieve complexiteit. De relatieve bijdrage van taakcomplexiteit en verbaalculturele aspecten aan geobserveerde crossculturele prestatieverschillen wordt vergeleken. Een steekproef van 1228 basisschool leerlingen, leeftijd 6 tot en met 12 jaar, is geselecteerd uit verschillende regio's in Nederland. (De data van de 6 en 7 jarige kinderen zijn als een leeftijdsgroep geanalyseerd.) De steekproef bestond uit een groep van Nederlandse autochtone kinderen ($\underline{n}$ = 747), en een groep tweede generatie allochtone leerlingen ($\underline{n}$ = 474). In beide culturele groepen zijn beide geslachten even sterk vertegenwoordigd per leeftijdgroep. De meerderheid van de deelnemers die getest zijn kwamen uit stedelijke gebieden waar migranten hoofdzakelijk woonachtig zijn.

In deze studie zijn geen eerste generatie allochtonen meegenomen. Eerste generatie migranten neigen naar een lager niveau aan kennis van de Nederlandse taal en cultuur vergeleken met de tweede generatie migranten. Het opnemen van eerste generatie migranten zou de invloed van verbaalculturele aspecten op test prestaties hebben versterkt. Het beperken van de studie tot tweede generatie kinderen zorgde ervoor dat alle bestudeerde kinderen een bekend (en over crossculturele groepen gelijk) aantal jaren

van Nederlandse opleiding hebben genoten. Daarnaast was er voldoende beheersing van de Nederlandse taal voor test administratie.

Echter, het is bekend dat substantiële verschillen in kennis van de Nederlandse taal tussen autochtone en allochtone leerlingen voortduurt gedurende de basisschool periode, zelfs bij tweede generatie kinderen. (Verhoeven, 2000) Een relatief groot aantal cultuur gereduceerde tests zijn afgenomen bij deze multiculturele steekproef van schoolkinderen in Nederland. Drie g maten zijn gebruikt. De eerste is gedefinieerd zoals door Jensen, de tweede is afgeleid van de Skill Theory (Fischer, 1980), en de derde maat is ontleend aan Carroll's (1993) hiërarchisch model van de structuur van cognitieve vaardigheden. De culturele lading van alle items van de tests is vastgesteld door derdejaars Psychologie studenten. De verbale lading van de tests is vastgesteld door de woorden van de tests te tellen. Al deze variabelen zijn factor geanalyseerd waardoor geaggregeerde g en c factoren naar voren kwamen. Deze factoren zijn gebruikt om allochtone - autochtone prestatie verschillen te voorspellen. De resultaten van deze studie suggereren dat prestatie verschillen tussen autochtone en allochtone leerlingen beter voorspeld worden door een cultuur factor (c) dan door g (taakcomplexiteit).

Op basis van de resultaten van deze studie en studies gedaan in het verleden, kunnen wij met overtuiging concluderen dat alle instrumenten cultureel beladen zijn, zij het in verschillende maten. De meting van een pure g is een misvatting. Als de test wordt afgenomen met als doel de intelligentie van een allochtone leerling te meten, zonder de kennis van de taal en cultuur van de meerderheidgroep te willen betrekken, moet de voorkeur uitgaan naar een cultuur gereduceerde test. Zelfs als de cultuur gereduceerde test niet vrij is van culturele lading zal dit de beste schatting van de cognitieve vaardigheid geven. De meting van cognitieve vaardigheden van leerlingen die recent in Nederland zijn komen wonen en die het schoolsysteem binnenstromen, dient slechts met cultuur gereduceerde tests plaats te vinden. Als echter toekomstig gedrag voorspelt dient te worden waarbij kennis van de autochtone taal en cultuur relevant is, dienen instrumenten, die meer cultureel geladen zijn, gebruikt te worden. De reden hiervoor is dat de test prestaties op deze tests sterkere correlaties vertonen met het criterium gedrag (school succes of succes op het werk) dan cultuur gereduceerde tests. Het is aan te raden deze tests in combinatie met cultuur gereduceerde tests af te nemen omdat de test resultaten dan beide het cognitieve potentieel en voortgang in 'het doen zoals de Nederlanders het doen' verschaffen. Uiteindelijk zal de combinatie tot beter onderbouwde besluiten en interventies leiden. Indien slechts cultuur geladen tests gebruikt zouden worden en een kandidaat een hoog cognitief potentieel vertoont, maar een laag niveau van kennis van de Nederlandse taal en cultuur paraat heeft, zal het cognitief potentieel onderschat worden.

Het tweede doel van dit project, dat in hoofdstuk 3 gepresenteerd wordt, was om construct-, methode- en item bias te detecteren in verschillende cognitieve maten die ontwikkeld zijn in Nederland, en om bepaalde antecedenten van bias te identificeren. Twaalf cultuur gereduceerde tests afgeleid van twee gestandaardiseerde intelligentie

batterijen en twee TAART subtests zijn bij een groot aantal proefpersonen afgenomen. Exploratieve factor analytische oplossingen van subtest scores van beide culturele groepen, zijn vergeleken om construct bias te meten. De analyse van methode bias op de allochtone groep, gaf een schatting van de invloed van non-cognitieve persoonskarakteristieken (bekendheid met de Nederlandse cultuur en met computers) op test prestaties. Item bias werd gemeten met logistische regressie en ANOVA. Effecten van construct- en item bias werden in sommige tests geconstateerd, en methode bias werd in bijna alle tests gevonden.

Evidentie van methode bias werd afgeleid uit verschillende bronnen. Ten eerste, was culturele familiariteit een relatief goede voorspeller van allochtone prestaties op de meer gekristalliseerde tests (CITO toetsen), maar was relatief onsuccesvol in het voorspellen van prestaties in minder gekristalliseerde toetsen. De mediaan van de absolute gestandaardiseerde regressie coëfficiënt was .19 voor culturele familiariteit. Ten tweede, voorspelde generatie status de allochtone prestaties op vele taken, vooral in de gekristalliseerde tests. Ten derde, gemiddelde GNP van het geboorteland van de leerling en dat van de ouders liet een significant verband zien met prestaties (beide allochtone en autochtone leerlingen werden beschouwd). Ook hier bleek de correlatie met de meer gekristalliseerde maten sterker te zijn. De invloed van culturele familiariteit op prestaties, en de modererende rol van de graad van kristallisatie van de taken, verschaft sterke evidentie voor de aanwezigheid van methode bias.

De construct- en item bias resultaten zijn in overeenstemming met algemene resultaten in dit onderzoeksveld. Indien onze onderzoeksresultaten gebaseerd zouden zijn geweest op slechts deze resultaten, waren we misschien tot de conclusie gekomen dat bias niet een grote rol speelt in mentale test prestaties van allochtone leerlingen. Het plaatje verandert als methode bias wordt onderzocht. Er is aangetoond dat non-cognitieve participantgerelateerde factoren significante voorspellers zijnwaren van allochtone prestaties. Deze resultaten zijn in overeenstemming met die van hoofdstuk 2 waar de $\underline{c}$ factor minstens zo belangrijk bleek te zijn als taakcomplexiteit in de verklaring van prestatie verschillen tussen autochtone en allochtone groepen. Een combinatie van de resultaten van hoofdstukken 2 en 3 leidt tot de conclusie dat waarschijnlijk een groot deel van de geobserveerde prestatieverschillen op cognitieve tests toegeschreven kan worden aan culturele factoren van de instrumenten die gebruikt zijn in multiculturele context in Nederland.

Als laatste, werden antecedenten geanalyseerd voor conditionele en niet-conditionele maten van item bias. Test karakteristieken (zoals item moeilijkheidsgraad, interne consistentie, $\underline{g}$ factor, en $\underline{c}$ factor) werden gebruikt om niet-conditionele en conditionele bias te voorspellen. Scores van de niet-conditionele factor werden redelijk goed voorspeld door test karakteristieken. ($\underline{R}^2 = .29$, $\underline{p} < .001$) Scores van de conditionele bias factor werden niet goed voorspeld. Conditionele en niet-conditionele bias statistieken zijn mogelijk gevoelig voor verschillende bronnen van vervorming. Om meer specifiek te zijn, conditionele technieken identificeren bronnen van error die

onopgemerkt zouden kunnen blijven bij het gebruik van niet-conditionele methodes. Simpele item- of test karakteristieken kunnen waarschijnlijk niet de meer subtiele mechanismen zoals verschuivingen in betekenis met score niveau indentiviceren. Dit verklaart mogelijk waarom conditionele bias moeilijker te voorspellen valt met de door ons gebruikte variabelen.

Samenvattend, twee bevindingen komen voort uit deze laatste studie. Ten eerste, de nadruk in de literatuur op structurele equivalentie en item bias detectie kan tot een onderschatting leiden van de invloed van culturele (methode) bias. De meer bekende en misschien wel betere methodes voor de analyse van construct- en item bias horen niet geïnterpreteerd te worden als een aanduiding dat methode bias niet belangrijk is. In tegendeel, vanuit een crosscultureel perspectief is het moeilijk te begrijpen waarom methodegerelateerde factoren zo lang verwaarloosd zijn gebleven. Om een belangrijke vorm van bias buiten beschouwing te laten kan een vervormd beeld geven over de validiteit van vergelijkingen tussen groepen. Slechts door een gebalanceerde behandeling van en de betrekking van verschillende bronnen van bias kunnen we inzicht verwerven in de ware aard van geobserveerde crossculturele overeenkomsten en verschillen.

This thesis addresses the suitability of cognitive tests in multicultural educational settings. A new cognitive test TAART (Tilburgse Allochtone en Autochtone Reactietijd Test) was developed and applied, along with other IQ measures, to a large multicultural sample of primary school children in The Netherlands. Spearman's Hypothesis was tested. Language and cultural influences on test scores were introduced in the testing of Spearman's Hypothesis. This was done to distinguish between the variance contributed by the latent trait (g) the test is intended to measure and the variance contributed by non-latent trait properties of the vehicle being used to measure the latent trait. Construct-, method- and item bias measures were estimated and compared for all the instruments used.

Michelle Helms-Lorenz was born in Johannesburg, South Africa on 17 May 1966. In 1987 she immigrated to The Netherlands. In 1992 she graduated in Cross Cultural and Health Psychology at the University of Tilburg. The project underlying this thesis commenced in 1993. During this research period her three sons were born.

TILBURG UNIVERSITY

ISBN 90-75001-37-1